Universidade de Aveiro Departamento de Electrónica, Telecomunicações e Informática,
2012

Elton Monteiro Rodrigues

Caracterização Multi-Escalar de Tráfego em Redes Protegidas

Universidade de Aveiro Departamento de Electrónica, Telecomunicações e Informática,
2012

Elton Monteiro Rodrigues

Caracterização Multi-Escalar de Tráfego em Redes Protegidas

"Não cruze os braços diante de uma dificuldade, pois o maior homem do mundo morreu de braços abertos."

— Bob Marley

Universidade de Aveiro Departamento de Electrónica, Telecomunicações e Informática,
2012

Elton Monteiro Rodrigues

Caracterização Multi-Escalar de Tráfego em Redes Protegidas

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requesitos necessários à obtenção do grau de Mestre em Engenharia de Computadores e Telemática, realizada sob a orientação científica de Paulo Salvador e António Nogueira, Professores Auxiliares do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro.

dedicatória Dedico esta dissertação à minha família.

o júri / the jury

presidente / president

Prof. Doutor André Ventura da Cruz Marnoto Zúquete
Professor Auxiliar da Universidade de Aveiro (por delegação da Reitora da Universidade de Aveiro)

vogais / examiners committee

Prof. Doutor Joel José Puga Coelho Rodrigue
Professor Auxiliar da Universidade da Beira Interior

Prof. Doutor Paulo Jorge Salvador Serra Ferreira
Professor Auxiliar da Universidade de Aveiro (orientador)

Prof. Doutor António Manuel Duarte Nogueira
Professor Auxiliar da Universidade de Aveiro (co-orientador)

agradecimentos / acknowledgements

Esta dissertação representa o auge de uma etapa académica, razão pela qual a lista de agradecimento se adivinha longa.

Primeiramente, gostaria de agradecer aos meus pais, Ilidio Rodrigues e Maria Monteiro, pela confiança, incentivo e pelo incansável e incondicional apoio dado ao longo de todo o percurso escolar, provando que nem sempre é preciso estar perto para se marcar a diferença.

Ao Prof. Doutor Paulo Salvador, meu orientador, pela atenção, disponibilidade, saber, conhecimento e por toda a orientação científica oferecida. Ao Prof. Doutor António Nogueira, co-orientador, pela acessibilidade, experiência.

Aos meus irmãos, Romilton Rodrigues, Davidson Rodrigues, Danilson Rodrigues, Jailson Rodrigues e Emiterio Rodrigues, pelo apoio dado e pela confiança depositada em mim.

Aos meus tios, Luis Fortes, Pedro Fortes e João Baptista.

Não poderia deixar de agradecer aos meus amigos e colegas de caminhada, Ulisses Gomes, Ineias Silva, Odracir Almeida e Hortencia lima acreditando que sem eles seria tudo mais difícil.

A todos aqueles que diretamente ou indiretamente contribuíram em todo o percurso que culminou com a presente dissertação. Um agradecimento muito especial ao Igor Fonseca, Alirio Boaventura, João Cruz, Oscar Gomes, Ricardo Sanches, Marcos Cruz, Ivo Petiz, João Rodrigues, Miguel Frade, Carlos Miguel e ao Eduardo Rocha.

Ao Instituto de Telecomunicações (pólo de Aveiro) pelas excelentes condições de trabalho oferecidas.

E a todos que não mencionei aqui, pela brevidade exigida, mas sempre tiveram uma palavra amiga.

Resumo

Atualmente, a Internet pode ser vista como uma mistura de diversos serviços e aplicações que correm sobre protocolos comuns. O aparecimento de inúmeras aplicações *Web* mudou o paradigma de interação dos utilizadores, colocando-os num papel mais ativo, permitindo aos utilizadores da Internet partilhar fotos, vídeos e muito mais. A análise do perfil de cada utilizador, tanto em redes *wired* como *wireless*, tornou-se muito interessante para tarefas como a otimização de recursos da rede, personalização de serviços e segurança.

Nesta dissertação pretende-se recolher um conjunto sistemático de capturas de tráfego correspondentes à utilização de diversas aplicações *Web* e efetuar a caraterização estatística do tráfego correspondente a cada aplicação em redes protegidas. O tráfego obtido (e as respetivas estatísticas) será posteriormente utilizado para validar metodologias de identificação de aplicações e caraterização do perfil de utilizadores da Internet. O desenvolvimento de diversas metodologias estatísticas permite caraterizar o tráfego associado a cada utilizador (tanto em redes *wireless* como *wired*) com base em informação estatística do tráfego por ele gerado enquanto utiliza os diversos serviços de rede. Neste sentido, é muito importante dispor de capturas de tráfego real que sejam representativas de uma utilização comum das diversas aplicações *Web*. Serviços *on-line* como notícias, *email*, redes sociais, partilha de fotografias e de vídeos podem ser estudados e caraterizados através da análise estatística do tráfego gerado pela utilização de aplicações como jornais *on-line*, Youtube, Flickr, GMail, Facebook, entre outras.

Ao extrair as métricas de tráfego ao nível da camada 2, realizar a decomposição baseada em Wavelets e analisar os escalogramas obtidos, será possível avaliar as diferentes componentes de tempo e de frequência do tráfego analisado. Será então possível definir um perfil de comunicação capaz de descrever o espetro de frequência característico de cada aplicação *web*. Consequentemente, será possível identificar as aplicações utilizadas pelos diferentes clientes ligados e criar perfis de utilizadores com precisão.

palavra-chave

Perfil de Utilizador, Tráfego Internet, Aplicações Internet, Rede Protegida, Perfil de Tráfego, Análise Multi-Escalar, Decomposição baseada em Wavelet, Escalograma.

Abstract

Nowadays, Internet can be seen as an mix of services and applications that run over common protocols. The emergence of several web-based applications changed the users interaction paradigm by placing them in a more active role, allowing users to share photos, videos and much more. The analysis of each user profile, both in wired and wireless networks, can become very interesting for tasks such as network resources optimization, service customization and security. This thesis aims to collect a systematic set of traffic captures corresponding to the use of several web-based applications in protected networks and perform a statistical traffic characterization for each application. The captured traffic (and the corresponding statistics) will be subsequently used to validate the methodologies developed to identify applications and characterize the traffic associated to each user. There are several statistical methodologies that allows the identification of users profiles (on both wireless and wired networks) based on statistical information collected from the traffic generated while using the different network services. In this sense, it is very important to have real traffic captures that are representative of a common use of several web-based applications. *On-line* services, such as news, *e-mail*, social networking, photo sharing and videos can be studied and characterized through the statistical analysis of the traffic captured while using applications such as *on-line* newspapers, Youtube, Flickr, GMail, Facebbok, among others. By extracting layer 2 traffic metrics, performing a wavelet decomposition and analyzing the obtained scalograms, it is possible to evaluate the time and frequency components of the analyzed traffic. A communication profile can then be defined in order to describe the frequency spectrum that is characteristic of each web-based application. By doing that, it will be possible to identify the different applications used by the connected clients and build accurate users profiles.

Key words

User Profiling, Internet Traffic, Internet Application, Protected Network, Traffic Profiling, Multi-Scale Analysis, Wavelet Decomposition, Scalogram.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

CoS	*Class-of-Services*
COI	*Community of Interest*
CWT	*Continuous Wavelet Transform*
DBSCAN	*Density Based Spatial Clustering of Applications With Noise*
DPI	*Deep Packet Inspection*
DWT	*Discrete Wavelet Transform*
FCA	*Formal Concept Analysis*
FP	Falso Positivo
FPGA	*Field Programmable Gate Array*
FFT	*Fast Fourier Transform*
FTP	*File Transfer Protocol*
GTK	*GIMP Toolkit*
GIMP	*GNU Image Manipulation Program*
HD	*High Definition*
HTTP	*Hypertext Transfer Protocol*
HTTPS	*HyperText Transfer Protocol Secure*
IANA	*Internet Assigned Numbers Authority*
IAT	*Inter-Arrival Time*
IC	Intervalo de Confiança
IMAP	*Internet Message Access Protocol*
IP	*Internet Protocol*
IPSec	*Internet Protocol Security*
ITU	*International Telecommunication Union*

ISP	*Internet Service Provider*
NN	*Nearest Neighbour*
k-NN	*k-Nearest Neighbors*
ML	*Machine Learning*
LDA	*Linear Discriminate Analysis*
Libpcap	*Library Packet Capture*
P2P	*Peer-to-Peer*
POP3	*Post Office Protocol 3*
QoS	*Quality-of-Service*
SP	*Service Providers*
SMTP	*Simple Mail Transfer Protocol*
SSH	*Secure Shell*
STD	*Standard Deviation*
TCP	*Transmission Control Protocol*
TF	Transformada de Fourier
TW	Transformada de Wavelet
UA	Universidade de Aveiro
UDP	*User Datagram Protocol*
VCD	Vídeos de Curta Duração
VLD	Vídeos de Longa Duração
VP	Verdadeiro Positivo
WEKA	*Waikato Environment for Knowledge Analysis*

Capítulo 1

Introdução

1.1 Enquadramento

Nos últimos anos assistiu-se a uma evolução extraordinária da Internet, quer em termos de volume de tráfego quer da variedade de aplicações disponibilizadas, tornando-se na mais poderosa plataforma de comunicação no planeta e constituindo um excelente meio para aceder e partilhar diversos tipos de informação, serviços e aplicações. Hoje em dia, os utilizadores de Internet podem assistir a vídeos *on-line*, usar aplicações de *chat*, fazer chamadas de voz e vídeo e muito mais. O aparecimento de aplicações em tempo-real de transmissão de voz e vídeo mudou de forma significativa o tipo de utilização da Internet. A Rede também começou a ser intensamente utilizada por uma ampla gama de pessoas para fins de investigação, comerciais e financeiros, proporcionando-lhes a possibilidade de realizar diversas operações financeiras usando serviços bancários *on-line*. Além disso, o recente aparecimento de serviços *Web* 2.0 mudou a própria Internet e a forma como os utilizadores interagem com ela. Atualmente, os conteúdos disponíveis na Internet são mais centrados no utilizador, uma vez que os utilizadores são agora produtores ativos de conteúdos e informação, sendo capazes de partilhar esses conteúdos *on-line* com a comunidade. Apesar das características promissoras e revolucionárias que estas aplicações apresentam, existem vários problemas de segurança e diversas vulnerabilidades [4].

Com a constante evolução da Internet, tanto em tamanho como em complexidade, o desafio de assegurar a sua gestão e segurança passa a estar intrinsecamente ligado à compreensão das aplicações e do tráfego por elas gerado. Atualmente, a satisfação das necessidades dos clientes é um requisito essencial e obrigatório e a capacidade de mapear o tráfego na correspondente aplicação oferece informações valiosas sobre os recursos a alocar para o tráfego, o que é essencial para garantir uma elevada disponibilidade dos serviços da rede.

Assim, nos últimos anos o desenvolvimento de métodos de caraterização de perfis de utilizadores e perfis de tráfego tornou-se numa tarefa de extrema importância para diversos fins, como a avaliação do desempenho da rede, gestão de recursos, personalização de serviços e segurança. Com base nesses métodos os *Provedores de Serviços de Internet* (*Internet Service Provider* (ISP)) e gestores de rede podem inferir mais facilmente a largura de banda, atribuir o tráfego conveniente à Classe de Serviço (*Class-of-Services* (CoS)) mais adequada, otimizando melhor os recursos de rede e distribuindo-os melhor pelos vários utilizadores. Podem assim ser alcançados melhores padrões de Qualidade de Serviço (*Quality-of-Service* (QoS)) para cada cliente ligado. Além disso, ao criar perfis de utilizadores com precisão, os administradores de

rede podem criar grupos de utilizadores que solicitam conteúdos semelhantes, facilitando a entrega de conteúdos e serviços adequados e relacionados. Desta forma, podem ser criados novos modelos de negócio, com o consequente aumento de receitas. No sentido de prever futuras necessidades dos utilizadores, é também possível alocar recursos de rede em tempo útil, evitando a sua futura saturação.

A personalização dos serviços e dos conteúdos também pode ser melhorada: a entrega de conteúdos e aplicações relacionadas pode ser facilitada a partir do conhecimento das aplicações e conteúdos que são mais solicitados pelos diversos utilizadores da rede.

Também podem ser obtidas melhorias significativas na segurança da rede: através do mapeamento preciso entre o tráfego e a aplicação de origem, é possível detetar fluxos gerados por aplicações ilícitas, fluxos que apresentam padrões suspeitos ou utilizadores que correspondem a perfis ilícitos, acionando alarmes e aplicando alguns métodos de defesa, tais como desligar utilizadores que representam perigo. Através da deteção oportuna de ataques de segurança ou *hosts* comprometidos, torna-se mais fácil conseguir uma melhor proteção dos clientes ligados e da infra-estrutura de rede, evitando assim perdas monetárias significativas. Desta forma, os clientes ligados podem usufruir de uma melhor qualidade de serviço e os administradores podem fazer um melhor uso da infra-estrutura de rede.

Muitas metodologias têm sido propostas para resolver o problema da classificação de tráfego, as quais tiveram que evoluir em conjunto com a complexidade das aplicações, serviços da Internet e da própria Internet em si [5]. A maioria das metodologias existentes baseia-se na análise estatística do tráfego da Internet ou na inspeção profunda do conteúdo dos pacotes (DPI). No entanto, restrições tais como a existência de tráfego cifrado e diversos aspetos legais impedem a implementação eficiente dessas abordagens de classificação. A principal razão por detrás destas restrições é a proteção da confidencialidade e privacidade das comunicações *on-line* dos utilizadores. Na verdade, a privacidade é um aspeto fundamental a ter em conta na análise de tráfego e têm sido propostas diversas abordagens para lidar com estas restrições [6]. São necessários novos paradigmas e metodologias de análise de tráfego de Internet no sentido de proporcionar mecanismos capazes de lidar com todas as restrições. A abordagem de classificação do tráfego proposta neste trabalho será validada através da análise do tráfego enviado para vários clientes ligados a uma rede *wired* e da posterior identificação das aplicações *web* que estão a ser utilizadas pelos diferentes clientes. Os resultados obtidos comprovam que é possível atribuir com precisão o tráfego à sua aplicação *web on-line* de origem, proporcionando uma descrição confiável e precisa sobre o uso de aplicações baseadas na *web*. O uso de métricas da camada 2 permite que esta abordagem de classificação se torne adequada à classificação de tráfego cifrado, em que o *payload* dos pacotes não está disponível, e também permite contornar restrições tecnológicas e legais que impedem a inspeção do conteúdo dos pacotes, garantindo uma proteção eficaz das suas redes e dos *hosts* das restantes redes.

1.2 Motivação

A questão da classificação de tráfego tem sido recentemente objeto de vários estudos. A necessidade de novas metodologias de classificação que possam lidar com a complexidade das redes existentes, com a sua crescente largura de banda e com o crescimento das várias aplicações de Internet foi uma das motivações para este trabalho. Outra motivação importante foi a necessidade de identificar com precisão tráfego ilícito e tráfego que apresenta padrões suspeitos. Para além disso, as metodologias de classificação existentes não realizam a inspeção do conteúdo do tráfego capturado, não lidando portanto com as várias restrições de privacidade impostas pelos ISPs e com a criptografia dos conteúdos, que impede a análise do tráfego capturado.

1.3 Objetivos

Este trabalho tem como objetivo principal o desenvolvimento de metodologias de classificação que permitam a identificação e a caraterização do tráfego gerado por aplicações *Web*. Tais tarefas são essenciais para a construção de perfis de utilizadores, uma questão de extrema importância na gestão de redes. As metodologias propostas também devem conseguir lidar com as diferentes limitações impostas pelas restrições de confidencialidade e integridade existentes, que impedem a inspeção dos dados transmitidos a partir do tráfego capturado.

Outro objetivo deste trabalho, não menos importante do que o acima citado, é o desenvolvimento de técnicas de *profiling* de utilizadores em redes *wired* e redes *wireless*. A partir da monitorização não promíscua de todas as máquinas ligadas, e através da medição das métricas do tráfego capturado, é possível construir o perfil de um utilizador que esteja a usar um conjunto de aplicações de Internet.

Em resumo, neste estudo pretende-se, a partir da recolha de um conjunto sistemático de capturas de tráfego correspondente à utilização de diversas aplicações *Web*, efetuar a caraterização estatística do tráfego correspondente a cada aplicação e construir o perfil dos utilizadores Internet que as estejam a usar. da Internet.

1.4 Estrutura da Dissertação

Após este capítulo introdutório, a dissertação apresenta a seguinte estrutura:

- No capítulo 2 é apresentado o estado da arte das metodologias de classificação de tráfego e caraterização do perfil de utilizadores da Internet, bem como os projectos e organizações existentes nesta área. É ainda efetuada uma avaliação do desenvolvimento atual de métodos de *profiling* de utilizadores na Internet.

- O capítulo 3 apresenta algumas definições relevantes que são comuns nas abordagens de classificação de tráfego e que incluem, por exemplo, a definição de *data-stream* como um conceito de agrupamento de tráfego, que representa o objecto de análise, modelação e classificação. É feita a apresentação das aplicações da Internet que serão estudadas, bem como os diferentes padrões de tráfego gerado, sendo ainda explicado o método de captura do tráfego. Posteriormente, será feita a comparação entre algumas das mais utilizadas metodologias de análise de sinal, juntamente com uma explicação das vantagens e inconveniências associadas a cada método de decomposição. São apresentadas e discutidas as transformações contínuas e discretas, destacando-se as vantagens e os problemas associados a cada tipo. Será ainda explicado como é efetuada a análise através da transformada de wavelet contínua ou discreta (*Continuous Wavelet Transform* (CWT) ou *Discrete Wavelet Transform* (DWT)). São também dadas algumas definições preliminares que serão utilizadas de forma intensa nos capítulos seguintes. Ainda no capítulo 3 é apresentado o laboratório, as configurações da máquina de testes e da rede em que foram realizados os diferentes testes, sendo explicada a forma como foi feita a recolha de dados das aplicações *web* analisadas, mostrados os procedimentos utilizados no processamento dos dados obtidos nos diferentes testes e ainda as diferentes análises feitas, a partir de todos os dados que se obtiveram para os diferentes testes realizados.

- O capítulo 4 centra-se na apresentação da metodologia de classificação. É descrita a arquitetura do sistema, os procedimentos e a estratégia adotada para a implementação do método proposto.

- No capítulo 5 são apresentados os resultados obtidos e que comprovam que a abordagem proposta é adequada tanto para a classificação de tráfego como para a criação de perfis de utilizadores.

- Por fim, o capítulo 6 é dedicado ás conclusões e a propostas de trabalho futuro baseado no sistema apresentado.

Capítulo 2

Estado de Arte

2.1 Introdução

Neste capítulo será inicialmente feita uma resenha do que tem vindo a ser desenvolvido na área da classificação de tráfego e caraterização do perfil de utilizadores da Internet, salientando no entanto que nesta dissertação se adotou uma definição muito específica de perfil do utilizador, mais orientada para o conjunto de aplicações *web* que cada utilizador utiliza e com as quais interage. Neste sentido, este trabalho difere de outros trabalhos mencionados pelo facto de definir/descrever um perfil de utilizador como um conjunto de aplicações utilizadas baseadas na *web*, particularmente aplicações que permitem aos utilizadores partilhar informações e conteúdos *on-line*.

2.2 Classificação de Tráfego e Caraterização do Perfil de Utilizadores da Internet

O problema da classificação do tráfego da Internet tem sido estudado desde há muitos anos e constitui um campo de pesquisa de extrema importância. De facto, a capacidade de associar com precisão o tráfego capturado com a sua aplicação *Web* de origem tem uma importância crucial para muitas atividades relacionadas com a utilização e gestão das redes de comunicação. Os administradores de redes podem facilmente criar e identificar as aplicações com maior tendência de uso, facilitando a gestão de várias tarefas da rede como a classificação de tráfego, otimização e personalização dos serviços da rede [7]. Por outro lado, a identificação das aplicações pode ser conseguida através do mapeamento preciso do tráfego da rede, ajudando o administrador da rede na identificação das aplicações que podem mudar os pedidos dos recursos da rede e, consequentemente, ajudando a lidar com a possível saturação da rede. A segurança da rede é também um dos aspetos que pode ser melhorado através do mapeamento preciso entre o tráfego e as aplicações de origem, uma vez que a identificação do tráfego ilícito ou tráfego que apresenta comportamento suspeito será facilitada. O objetivo principal da classificação de tráfego prende-se com a capacidade de inferir, através das propriedades estatísticas do tráfego capturado, caraterísticas que permitam associar com precisão o tráfego à sua aplicação Internet original. Várias técnicas têm sido propostas no sentido de aumentar o desempenho das metodologias de classificação de tráfego. Na subseção seguinte são apresentadas algumas metodologias de classificação já propostas na literatura, enumerando as suas principais vantagens e as consequências que tiveram.

2.2.1 Metodologia de classificação baseada em números de porto

Historicamente, uma das formas mais comuns de classificação de tráfego tem sido a classificação baseada no número de porto, que faz uso dos números dos portos usados pelas aplicações na camada de transporte. Portanto, a classificação do tráfego pode ser efetuada através de uma simples associação entre o número do porto usado e os serviços correspondentes. Os portos são divididos em três classes: portos bem conhecidos (*Well Known Ports*), portos registados, e portos dinâmicos e/ou portos privados. A classe dos portos bem conhecidos abrange os portos entre 0 a 1023 e corresponde aos portos reservados para uso privilegiado, enquanto os portos registados são aqueles que se situam entre 1024-49151. Finalmente, os portos dinâmicos e/ou privados são os que se situam entre 49152 e 65535. Esses números são atribuídos pela *Internet Assigned Numbers Authority* (IANA) [8]. Muitas aplicações utilizam os *Well Known Ports* nas suas máquinas locais como um ponto de encontro para que outras máquinas possam iniciar a comunicação.

Alguns trabalhos usando esta abordagem foram propostos na literatura, tendo sido capazes de alcançar resultados precisos para alguns cenários [9].

No entanto, esta abordagem tem limitações. Em primeiro lugar, algumas aplicações podem não ter os seus portos registados em IANA, caso das aplicações *Peer-to-Peer* (P2P) como o Napster e Kazaa [10]. Além disso, muitas aplicações modernas usam portos dinâmicos, o que faz com que a técnica de classificação baseada em portos se torne ineficiente, com taxas de precisão situadas entre 30% e 70% [11].

Um estudo realizado por Manzano et. al. [12] provou que a análise baseada no número do porto já não fornecia resultados precisos quando comparada com outros métodos de identificação, chegando a resultados que apontavam para valores de tráfego desconhecido (não classificado) entre 40% a 65%. Este estudo também confirmou que o tráfego desconhecido foi mais evidente em períodos noturnos, permitindo chegar à conclusão que o tráfego foi gerado por aplicações P2P.

Em [13], os autores desenvolveram várias heurísticas de classificação que lhes permitiam identificar tráfego P2P sobre portos *nonstandard*. Os autores concluíram que, de acordo com o protocolo e a métrica que foi utilizada, cerca de 30% a 70% do tráfego P2P não poderia ser identificada usando portos *standard*.

Os autores de [14] tentaram ultrapassar a limitação da classificação do tráfego baseada em portos conhecidos desenvolvendo heurísticas capazes de medir o tráfegoP2P oculto. O trabalho consistia em *reverse engineering* dos protocolos e na identificação de sequências de carateres característicos transportados no *payload*. Os resultados obtidos mostraram que as aplicações P2P evoluíram na utilização de números de porto arbitrários para a comunicação.

Moore e Papagiannaki [15] mostraram que usando esta técnica mais de 28% do tráfego capturado não poderia ser classificado. De facto, em algumas circunstâncias a aplicação de técnicas criptográficas ao nível da camada *Internet Protocol* (IP) poderia ofuscar o cabeçalho *Transmission Control Protocol* (TCP) ou *User Datagram Protocol* (UDP), o que tornaria impossível saber os números reais dos portos.

2.2.2 Metodologia de classificação baseada na análise do *payload* dos pacotes (DPI)

A análise baseada em pacotes consiste na captura e análise do seu cabeçalho. Várias informações são obtidas a partir do cabeçalho do pacote, tal como o endereço IP de origem (*srcIP*), o endereço IP de destino (*dstIP*), o porto de origem (*srcPrt*), o porto de destino (*dstPrt*) e o número do protocolo. Nesta abordagem o *payload* dos pacotes capturados é analisado com o propósito de encontrar assinaturas digitais (sequências de *bytes/strings* específicos) precisas que podem ser usadas para identificar a aplicação Internet [12] que lhes deu origem. Diversos estudos mostram que essas abordagens são eficientes para a classificação do tráfego Internet, incluindo tráfego P2P. Várias técnicas baseadas na análise do *payload* foram propostas na literatura [[15], [11], [16], [17]].

Numa das primeiras obras, apresentada em [11], os autores propuseram assinaturas ao nível das aplicações para uma identificação precisa do tráfego P2P. Os autores analisaram a documentação e os *traces* pertencentes a diferentes clientes P2P com o objetivo de obterem assinaturas para cada uma das aplicações, que foram posteriormente usadas para desenvolver filtros que podiam controlar eficientemente o tráfego P2P em *links* de alta velocidade. Os autores alcançaram resultados de classificação com uma alta taxa de precisão, com menos de 5% falsos positivos e de falsos negativos. Entretanto a abordagem exigia um conhecimento prévio de cada aplicação, no sentido de desenvolver as assinaturas certas para cada aplicação, o que impedia a adaptação automatica desta abordagem a novas aplicações.

Um outro trabalho importante [14] focou-se nos relatórios que reivindicaram uma redução significativa no tráfego de partilha de ficheiros em aplicações P2P. Os autores começaram por medir o tráfego de todos os protocolos P2P e, usando *reverse engineering*, conseguiram analisar estes protocolos, identificando *strings* característicos no seu *payload*. Várias Heurísticas de classificação foram propostas para determinar com precisão se o tráfego analisado foi gerado ou não por um protocolo P2P. Estas heurísticas focaram-se na análise dos portos de origem e destino para determinar se correspondem ao porto P2P respetivo, caso o fluxo fosse marcado como P2P. Posteriormente, os autores compararam o *payload* de cada pacote com as assinaturas características obtidas, o que lhes permitiu identificar com precisão o protocolo P2P. Mas os resultados mostraram-se contraditórios, revelando uma diminuição no volume de tráfego P2P e apontando também alguns obstáculos na identificação precisa deste tipo de tráfego.

Em [15], os autores utilizaram a análise do *payload* para quantificar os erros associados à abordagem de classificação baseada nos portos. O tráfego utilizado foi capturado da partir de um *site* citado como *Genome Campus*, hospedando várias instalações de biologia. O tráfego foi agrupado em fluxos para um processamento mais eficiente das informações recolhidas e para a recuperação do contexto de identificação da aplicação de rede que corresponde a cada fluxo. Os autores obtiveram resultados muitos precisos.

Haffner et al. [17] propuseram em 2005 uma nova abordagem para a construção de assinaturas de aplicações utilizando a técnica *Machine Learning* (ML). Diferente dos outros trabalhos, esta abordagem faz uso dos primeiros n bytes de um fluxo de dados. Os autores exploraram a extração automática de assinaturas de aplicações através da aplicação de três

algoritmos baseados em análise estatística, nomeadamente Naive Bayes, AdaBoost e modelo *Regularized Maximum Entropy*. Desta maneira, a abordagem de classificação proposta foi capaz de construir assinaturas de aplicações para uma grande gama de diferentes aplicações de rede. As aplicações estudadas foram *File Transfer Protocol* (FTP), *Simple Mail Transfer Protocol* (SMTP), *Post Office Protocol 3* (POP3), *Internet Message Access Protocol* (IMAP), *Hypertext Transfer Protocol* (HTTP), *HyperText Transfer Protocol Secure* (HTTPS) e *Secure Shell* (SSH), tendo sido escolhidas por cobrirem uma ampla gama de classes de aplicações. Neste caso, foi muito fácil obter um conjunto de *traces* de treino devidamente certificados, visto que estas aplicações usam os seus portos *standard*. Os autores também analisaram a durabilidade das assinaturas extraídas por meio da classificação dos *traces* de tráfego recolhidos sete meses após o primeiro conjunto de dados. Os erros de classificação apenas aumentaram ligeiramente, o que indicou que os classificadores mantinham um bom desempenho e as assinaturas poderiam ser usadas para períodos de tempo mais longos.

Em [12], os autores focaram os seus estudos em medição do tráfego de aplicações P2P na Internet, tendo apenas analisado as *flags* TCP SYN, FIN e RST, com o objetivo de obterem informações ao nível do transporte de tráfego P2P. Estes autores mostraram que a abordagem baseada no porto é incapaz de classificar 30% a 70% do tráfego capturado. Está técnica apresenta várias desvantagens: tráfego com comportamento desconhecido não pode ser classificado; quando o tráfego é transportado através de um túnel seguro, os números dos portos TCP e as *flags* TCP podem não estar disponíveis, não sendo possível efetuar a classificação do tráfego.

Apesar da técnica de classificação baseada no *payload* evitar dependências de números de portos fixos, apresenta muitas desvantagens. Primeiro, essas técnicas apenas identificam o tráfego cujas assinaturas estão disponíveis e são incapazes de classificar qualquer outro tráfego. Segundo, a análise da *payload* é muito complexa e requer um grande poder computacional na identificação do tráfego, uma vez que é feita uma análise completa do *payload*. Finalmente, as leis de privacidade não permitem que os administradores inspecionem o *payload*, sendo impossível uma análise direta da sessão e dos conteúdos da camada de aplicação. Consequentemente, está técnica falhará se o *payload* for cifrado.

2.2.3 Metodologia de classificação baseada na análise estatísticas dos pacotes

Esta abordagem trata do problema da classificação da aplicação como um problema de estatística. O conceito principal desta abordagem é que o tráfego gerado pelo mesmo protocolo irá apresentar o mesmo perfil. Surge como uma solução que permite superar as restrições apresentadas pelas abordagens anteriores, uma vez que apenas os cabeçalhos dos pacotes são analisados [18]. A classificação do tráfego da rede é baseada em várias métricas estatísticas do fluxo de pacotes, como por exemplo o número de pacotes, tamanho do pacote, o tempo de chegada, etc. O método mais discutido é a técnica de análise Bayesiana [19], [20], [21].

Em [19] vários discriminadores de fluxos foram propostos, tendo sido utilizadas técnicas ML para selecionar os melhores discriminadores de classificação. Os discriminadores utilizados para esta análise incluíram portos TCP, o *Inter-Arrival Time* (IAT) e a sua Transformada de Fourier (TF), a carga e a largura de banda efetiva. Ao realizar alguns refinamentos so-

bre o classificador, os autores foram capazes de chegar a uma precisão de 95%. A principal desvantagem desta abordagem é que requer muitos *traces* de treino, uma vez que a relação entre os *traces* de treino e de teste é de 1:1, o que nem sempre é possível de conseguir na prática. Além disso, se houver alguma alteração nos parâmetros da rede ou do tráfego, os classificadores têm de voltar a ser treinados.

Em [22], os autores criaram perfis de comportamento que descrevem padrões dominantes das aplicações estudadas e os resultados e classificação obtidos mostraram que esta abordagem era bastante promissora.

Uma obra muito importante para classificar aplicações P2P foi proposta em [23]. A identificação das aplicações P2P era feita com base nos padrões de ligação. Esta técnica utilizava duas heurísticas de classificação. A primeira examinava os pares IP de origem e destino, usando os protocolos TCP e UDP para a transferência de dados. A segunda heurística era baseada na forma como os pares são ligados entre si, e os autores examinaram todas os pares *srcIP, srcPort destIP e destPort*. Os pares em que o número de IPs ligados é igual ao número de portos ligados são definidos como P2P. Este metodologia foi capaz de identificar mais de 90% dos bytes P2P, mesmo com taxas de bits tão elevadas como 220 Mbp/s.

Noutro trabalho inovador [24], os autores apresentaram uma metodologia para associar o tráfego capturado e a sua CoS. Os autores basearam-se em dois métodos de classificação, *Nearest Neighbour* (NN) e *Linear Discriminate Analysis* (LDA), no sentido de mapear com sucesso as aplicações nas diferentes classes de QoS utilizando até quatro atributos. O número de classes considerado era pequeno, sendo as classes conhecidas *a priori*. Os atributos utilizados na classificação foram agregados em períodos de mais de 24 horas. Foram utilizadas grandes quantidades de *traces* de tráfego de diferentes locais da rede para avaliar a precisão da metodologia, tendo sido obtidas taxas de erro baixas.

Blinc [25] propôs uma nova abordagem, baseada no padrão de comportamento dos *hosts* na camada de transporte, para classificar fluxos de tráfego de acordo com o tipo de aplicação. Os padrões são analisados em três níveis diferentes. Em primeiro lugar, o nível social, que analisa a popularidade do *host*. Neste nível é investigado o comportamento do *host* em relação ás suas comunicações com outros *hosts*. Em segundo lugar, o nível funcional, que investiga o que o *host* faz. Neste nível é analisado se o *host* pretendido fornece ou consome o serviço. Finalmente, o nível de aplicação, que se destina a identificar aplicações e suas origens. Os autores afirmaram que a sua abordagem classificava entre 80% e 90% do número total dos fluxos em cada *trace* com uma precisão de mais de 95%, além de detetar aplicações maliciosas e desconhecidas. No entanto, este método não pode ser aplicado a tráfego *on-line* em tempo real, devido à limitação da velocidade de classificação, sendo mais adequado para análise de tráfego *off-line*.

A técnica proposta em [26] é baseada na análise do comportamento dos computadores: é realizada a análise do fluxo através da observação dos cinco primeiros pacotes de uma ligação TCP para identificar a aplicação. Ao contrário de outras técnicas, onde a classificação das aplicações ocorre somente após o final do fluxo TCP, esta técnica utiliza apenas o tamanho dos primeiros pacotes de uma ligação para classificar o tráfego. A ideia da classificação *on-the-fly*, denominação dada pelos autores, é identificar com precisão a aplicação associada a

um fluxo TCP o mais cedo possível. Essa técnica utiliza o conceito de *clusters* não supervisionados para descobrir grupos de fluxos que partilham um comportamento de comunicação comum. A classificação é dividida em duas fases, a primeira fase, denominada de fase de treino, é utilizada para aprender e detetar comportamentos comuns através de um conjunto de dados de treino, com o objetivo de criar classes de comportamentos. A segunda fase é designada de fase de classificação, e é utilizada para determinar a aplicação associada a cada fluxo TCP. Esta técnica obtém mais informação do que a técnica DPI, respeitando os aspetos de privacidade e as restrições existentes. Os resultados obtidos indicam que as aplicações TCP bem conhecidas foram identificadas com uma taxa de precisão superior a 80%. Entretanto, a técnica apresenta vários problemas, como o facto de os pacotes que chegam fora de ordem causarem alterações na representação espacial dos fluxos de tráfego, o que irá afetar a qualidade da classificação. Além disso, aplicações com comportamento inicial semelhante poderão ser classificadas com o mesmo rótulo. Por outro lado, esta abordagem não lida com criptografia de tráfego, que pode impedir a análise dos cabeçalhos do TCP.

No trabalho publicado em [27] os autores demonstraram que a análise de *cluster* pode ser usada para identificar grupos de tráfego que são semelhantes, recorrendo apenas a estatísticas da camada de transporte. Três algoritmos de *clustering*, K-Means, *Density Based Spatial Clustering of Applications With Noise* (DBSCAN) e *AutoClass*, foram utilizados para resolver o problema da classificação de tráfego de rede. O algoritmo K-Means produz *clusters* com forma esférica, enquanto o algoritmo DBSCAN tem a capacidade de produzir *clusters* não esféricos. As diferentes formas dos *clusters* obtidos através do DBSCAN permitem encontrar um melhor conjunto de *clusters* que minimizam a quantidade de análises requeridas. O algoritmo *AutoClass* utiliza uma abordagem Bayesiana e pode determinar automaticamente o número de *clusters*. O algoritmo DBSCAN apresentou a melhor precisão de classificação, enquanto o K-Means foi a abordagem mais rápida.

Hu et al [28] propuseram uma abordagem baseada em perfis para identificar os fluxos de tráfego pertencentes a uma determinada aplicação P2P. Dependendo dos padrões dominantes da aplicação, foram construídos perfis de comportamento da aplicação de destino. Com base nesse perfil comportamental, foi usado um método com dois níveis para identificar o tráfego. No primeiro nível, determina-se se um *host* participa na aplicação de destino, comparando o seu comportamento com os perfis. No segundo nível, comparara-se cada fluxo do *host* com os padrões dos perfis de aplicação para determinar a que aplicação pertence esse fluxo. Os resultados mostraram que as aplicações P2P populares podem ser identificadas com uma precisão elevada. No entanto, o número de regras necessárias para a classificação é muito elevado, o que levanta algumas questões sobre a escalabilidade da abordagem, bem como a sua capacidade para classificar tráfego *on-the-fly*.

Num trabalho recente [29], os autores analisaram uma nova aplicação do algoritmo de *clustering* k-means conhecida como classificação *two way*. A classificação *two way* analisa um fluxo bidirecional como dois fluxos unidirecionais, e os autores mostraram, através de testes sobre tráfego real, que a sua abordagem aumenta a precisão de classificação em mais de 18% quando comparada com outras abordagens semelhantes. Esta precisão é alcançada através da geração de *clusters* menores, isto é, são necessárias menos comparações para classificar um fluxo. Este método oferece novas formas de melhorar a precisão e eficiência dos classificadores estatísticos ML, mantendo os tempos de treino rápidos associados ao k-means. No entanto,

este método sofre do facto de tráfego com o mesmo comportamento estatístico poder ser classificado como pertencendo à mesma aplicação, o que pode não ser verdade, e o tráfego com o comportamento desconhecido não ser classificado.

Técnicas de *clustering* são ferramentas úteis para o agrupamento de tráfego com características semelhantes [27], mas têm que confiar em outras técnicas de identificação para rotular os *clusters*. Classificadores ML também são baseados na análise estatística do tráfego da Internet e demonstraram uma precisão elevada na classificação do tráfego [21].

Obras mais recentes [[30], [31]], descrevem um novo método de deteção de anomalias de rede baseada em Transformada de Wavelet (TW). Modelos matemáticos, especialmente baseados em transformadas de *wavelet*, têm sido empregues na deteção de anomalias porque capturam correlações temporais complexas através de múltiplas escalas de tempo e encontram variações no comportamento do tráfego de rede. *Wavelets* são funções matemáticas que dividem os dados (sinais) em diferentes componentes de acordo com uma escala de interesse, permitindo realizar análises locais numa área específica do sinal. No entanto, estes trabalhos não exploraram as características multi-escala do tráfego de rede na deteção de anomalias.

2.2.4 Metodologia de classificação baseada na análise do tráfego em tempo real

Classificação de tráfego em tempo real (*Real-Time*) é de fundamental importância para certas operações de rede, tarefas de gestão, de estudo e de planeamento. Serve como entrada para sistemas de deteção de intrusões, é capaz de identificar a CoS para um mapeamento eficaz [10] no nível de QoS apropriado e também fornece estatísticas para monitorização da rede. Esta abordagem cria condições para que os ISPs melhorem a sua infra-estrutura tecnológica e o planeamento dos serviços de acesso à Internet. Os operadores da rede têm que ter conhecimento do tráfego que está a fluir para que possam reagir rapidamente no sentido de suportar as metas de comunicação das empresas [18].

Em 2006, Nguyen e Armitage [32] propuseram um método de classificação baseado nos primeiros N pacotes de um fluxo - denominado de classificação *sliding window*. A utilização de um pequeno número de pacotes para a classificação garante a correção da classificação e reduz o espaço de memória necessário para armazenar a informação dos pacotes no processo de classificação. Esta abordagem permite que a classificação se inicie em qualquer ponto do tempo. Esta técnica permite uma monitorização do fluxo do tráfego durante a sua vida, apresentando no entanto algumas limitações de recursos físicos. O trabalho propõe classificadores de ML baseados em vários sub-fluxos, extraídos dos fluxos originais das aplicações de Internet. A técnica foi posteriormente avaliada com a utilização do classificador Naive Bayes, obtendo altas taxas de precisão na classificação.

Num estudo recente, Bonfiglio et al. [33] utilizaram duas técnicas para identificar tráfego Skype em tempo real. A primeira técnica, baseada em testes *Pearsons Chi-Square*, detecta as impressões digitais do tráfego Skype através da análise do conteúdo das mensagens. Na segunda, baseada no teorema de Naive Bayes, o tráfego Skypes é detetado através do tamanho da mensagem e das características da taxa de chegadas. A precisão dos resultados obtidos foi verificada através da sua comparação com os resultados obtidos por DPI. Os

autores mostraram que a técnica Naive Bayes pode ser eficaz na identificação de tráfego de voz através do IP, independentemente da aplicação de origem, e que a técnica *Pearsons Chi-Square* é eficiente na identificação de tráfego Skype e de tráfego TCP cifrado. Note-se que a inspeção dos fluxos é dificultada pela utilização de técnicas criptográficas. Ao utilizar as metodologias de identificação apresentadas, a percentagem de falsos positivos caiu para quase zero.

Em [34], os autores tentaram descrever comportamentos de negociação através da captura de discriminadores de tráfego disponíveis nas primeiras fases de negociação de fluxos de rede, tendo implementado vários algoritmos ML para avaliar a precisão da classificação. Ao usar tais discriminadores, os autores concluíram que a abordagem proposta é adequada para a identificação de aplicações em tempo real.

A ferramenta *Waikato Environment for Knowledge Analysis* (WEKA) [35] faz uso de um conjunto de algoritmos de ML para avaliar a precisão de classificadores como Bayesian Networks, Multilayer Perceptron Network, Naive Bayes Tree, C4.5 Decision Tree, e Naive Bayes. Os autores mostraram que este trabalho aumentou a precisão em cerca de 8% a 21% quando comparado com alguns trabalhos anteriores. No entanto, este trabalho carece de novas metodologias de classificação, já que os autores apenas utilizaram algoritmos ML implementados no WEKA e a sua abordagem não é capaz de analisar tráfego cifrado.

Num trabalho mais recente [36], arquiteturas baseadas em *Field Programmable Gate Array* (FPGA) são utilizadas para acelerar a identificação com base em *k-Nearest Neighbors* (k-NN). De acordo com os resultados apresentados, foi obtida uma alta taxa de precisão (acima de 99%) na classificação de três aplicações multimédia.

Em [37], os autores propuseram uma técnica baseada em ML para identificar tráfego BitTorrent. Utilizaram um método de treino para a classificação dos sub-fluxos onde apenas uma parte do fluxo (tipicamente algumas centenas de pacotes) é utilizadao. A vantagem de utilizar sub-fluxos não é apenas no sentido de obter uma maior rapidez na classificação, mas também ter a capacidade de classificar corretamente uma classe com sucessivos sub-fluxos, mesmo se um sub-fluxo não puder ser classificado corretamente de forma individual. Foi obtida uma precisão de classificação a rondar os 98%. Apesar dos bons resultados obtidos na identificação de classes de tráfego para as versões específicas da aplicação, a técnica obteve resultados desfavoráveis na identificação de diferentes versões da mesma aplicação. Os autores provaram que os parâmetros estatísticos do tráfego cifrado e não cifrado produzidos pela mesma aplicação são semelhantes e, portanto, o *payload* cifrado não influecia os resultados de treino ou de classificação.

A classificação de tráfego de rede é uma atividade essencial para diversos sistemas, tais como sistemas de detecção de anomalias, gestão de rede e de QoS. As técnicas clássicas de classificação de tráfego envolvem normalmente a inspeção do conteúdo de pacotes e a análise dos portos TCP/UDP. Entretanto, a eficácia de tais técnicas têm-se mostrado bastante limitada como consequência da utilização de criptografia e do uso de portos não convencionais. Por outro lado, a classificação de tráfego baseada em métodos estatísticos e técnicas de ML tem suscitado um grande interesse, em virtude da sua capacidade de utilizar informações retiradas somente dos cabeçalhos dos pacotes, o que permite classificar o tráfego desconhecido sem a necessidade de analisar o conteúdo do pacote em detalhe para determinar o tipo de

aplicação gerada. Isto também significa que é necessário menos poder de processamento para executar a análise.

2.3 Profiling de Utilizadores

O *profiling* de utilizadores da Internet visa o desenvolvimento de técnicas de classificação que permitam criar perfis precisos dos diferentes (tipos de) utilizadores. Existem várias definições para perfil de utilizador [38], mas uma definição comum afirma que um perfil de utilizador consiste numa descrição dos seus interesses, comportamentos e preferências [39]. Portanto, consiste numa coleção de dados relacionados com o utilizador, que é adequada para o sistema que se pretende criar [40]. Os perfis de utilizadores podem ser construídos porque cada utilizador difere nas suas preferências, interesses, origens e objetivos [41].

Segundo a *International Telecommunication Union* (ITU), nos últimos dez anos o número de utilizadores da Internet quadruplicou (ver figura 2.1). Novos sistemas, aplicações e serviços são desenvolvidos a um ritmo elevado, aproveitando o potencial de um enorme conjunto de redes que interligam uma parte substancial da população mundial. A Internet adquiriu uma importância inquestionável na vida de pessoas e instituições, tornando-se num dos principais meios de comunicação, transferência de dados e consulta de informação.

Entender as características do comportamento dos utilizadores é uma tarefa que pode melhorar a qualidade de serviço do ambiente criado pela Internet e, além disso, contribuir para o desenvolvimento e evolução das aplicações utilizadas. A caraterização dos utilizadores pode facilitar o entendimento da interação entre utilizadores e provedores de serviços, bem como ajudar no projeto de sistemas com melhores métricas de qualidade de serviço, tais como, desempenho, disponibilidade de acesso, segurança e custo.

Com o crescimento da quantidade de dados e utilizadores *on-line*, vários trabalhos se têm centrado no desenvolvimento de técnicas de *profiling* de utilizadores.

A maior parte dos trabalhos foram realizados na área de deteção de anomalias no computador, nomeadamente na deteção de fraudes em redes de telecomunicações [[42], [43]]. As técnicas utilizadas vêm da área da análise estatística [[40], [44]] , *clustering* [45], regras Bayesianas [42], ou classificação baseada em redes neurais [46]. Combinações de métodos são também utilizadas [47].

Muitos trabalhos, como exemplo [39], têm abordado a questão de criar perfis de utilizadores com precisão, descrevendo as características mais importantes, embora o conjunto de características e consequentemente a definição do perfil do utilizador varie de acordo com o objetivo da classificação.

Em [40], os autores apresentam um método para validar o conjunto de regras usadas para a construção de perfis de utilizadores. Os autores propuseram um processo de construção de perfis de utilizadores que usa algoritmos de mineração de dados para descobrir associação e regras de classificação. Apesar de atingir regras eficientes para construção de perfis, os autores necessitam sempre da validação humana, embora pudessem ser implementadas diversas

abordagens para validar tais regras de forma eficiente.

Em [48] os autores construíram perfis *end-host* com o objetivo de se defenderem de ataques de *worms*. Um perfil é definido como uma comunidade de *hosts* através da qual interagem alguns sistemas, o que é definido como uma *Community of Interest* (COI). Os autores exploraram algumas propriedades das redes das empresas, tal como o conhecimento de todos os seus *hosts* e equipamentos, que permitem o controlo de todas as configurações em todos os *routers* e *switches* da rede. Foram recolhidos dados de treino ao longo de várias semanas para obter um "padrão de comunicação normal". Os autores construíram então perfis capazes de tolerar algum desvio em relação ao perfil normal, no sentido de lidar com as mudanças na rede e dos próprios perfis dos utilizadores. Finalmente, alguns comportamentos conhecidos de *worms* foram utilizados para assegurar que os dados de treino estavam livres de ataques. Através das regras e dos perfis criados, os autores foram capazes de impedir ataques às redes monitorizadas. No entanto, a precisão desta abordagem depende dos perfis utilizados e presume-se que os padrões de um ataque se desviam sempre dos perfis de comunicação normal, o que nem sempre é verdade.

Em [49] foi proposta uma abordagem inovadora para a construção de perfis de utilizadores de redes por pesquisa personalizada. Os autores definiram e modelaram um perfil de utilizador como uma estrutura de redes de conceitos, permitindo o uso da teoria de *Formal Concept Analysis* (FCA). Um conceito contém a intenção de consultar um utilizador e reflete as preferências deste. Sempre que uma nova consulta é emitida, uma sessão de conceito de interesse é gerada e novos conceitos aparecem na rede de conceito atual, ou seja, um perfil de utilizador. Semelhanças entre os novos conceitos e os já existentes são também calculadas, sendo utilizada uma hierarquia de conceitos de referência para esta finalidade. Os resultados obtidos mostram que a abordagem proposta é capaz de melhorar a precisão dos resultados da pesquisa em termos de preferência pessoal.

Estão atualmente disponíveis muitas ferramentas de gestão de rede. Um exemplo é o Sistema de Gestão de Rede Aberta (Open SMN), uma plataforma de código aberto que executa muitas tarefas de gestão de rede [50], tais como eventos e gestão de notificações, garantia de serviço e avaliação de desempenho. A sua escalabilidade permite monitorizar milhares de dispositivos numa única rede. No entanto, a monitorização do serviço de *hosts* é baseada nos portos que o administrador associa com as diferentes aplicações de rede, o que constitui uma lacuna desta plataforma.

Em [5], a crescente complexidade da Internet, do número de utilizadores e dos serviços são discutidos para alcançar soluções de gestão flexível e escalável. Algumas orientações importantes e instruções de pesquisa são fornecidas para lidar com a crescente complexidade da Internet e das suas aplicações.

Um trabalho semelhante foi levado a cabo em [[51], [52]], no qual os autores afirmam que toda a informação necessária para construir o perfil de qualquer terminal da Internet está disponível à nossa volta. Portanto, para criar perfis de utilizadores precisos, os autores concluíram que é necessário efetuar uma consulta no motor de busca mais utilizado (Google) e dividir os resultados em várias *tags* que descrevem os serviços solicitados. Os resultados obtidos comprovaram que a abordagem é adequada para o objetivo proposto, possibilitando resultados ainda mais precisos do que algumas das ferramentas descritas anteriormente.

Num trabalho recente [53], através da análise do tráfego enviado para vários clientes ligados a uma rede *wireless*, o autor identifica as aplicações que estão a ser utilizadas por diferentes utilizadores. Os resultados obtidos comprovam que é possível atribuir com precisão o tráfego à sua aplicação *web on-line* de origem, proporcionando uma descrição confiável e precisa sobre o uso de aplicações baseadas na *web*. O uso de métricas da camada 2 permite que esta abordagem de classificação seja adequada para a classificação de tráfego cifrado, onde os conteúdos dos pacotes não estão disponíveis, e também permite contornar restrições tecnológicas e legais que impedem a inspeção do conteúdo dos pacotes, garantindo uma proteção eficaz das suas redes e dos *hosts* das restantes redes.

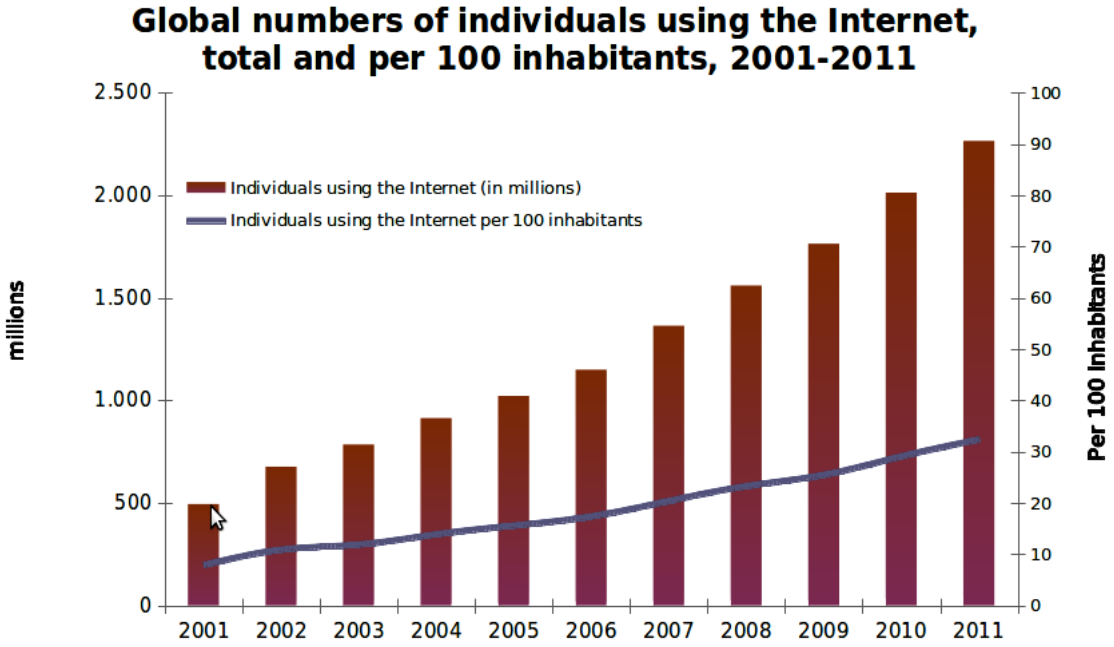


Figura 2.1: O acesso à Internet nos últimos dez anos [retirado de ITU *World Telecommunication*/ICT Indicators database [1]].

2.4 Sumário

Nesta secção foram apresentados os trabalhos mais relevantes das diferentes áreas relacionadas com a classificação: classificação de tráfego, deteção de ataques de segurança e *profiling* de utilizadores.

Capítulo 3

Noções Preliminares

Neste capítulo são apresentados vários conceitos importantes. O capítulo começa por fazer uma análise de diversas aplicações da Internet, identificando os diferentes mecanismos que geram e moldam o tráfego de cada uma das aplicações e descrevendo as suas principais características. Posteriormente, é apresentada a definição de *data-stream*, referindo a sua importância na análise das dinâmicas do tráfego. Tendo em consideração que estas dinâmicas estão espalhadas pelas diferentes sessões estabelecidas com *hosts* e servidores remotos, o conceito de *data-stream* deve ser suficientemente abrangente para conseguir incorporá-las. Seguidamente, é apresentado o tráfego capturado e as aplicações estudadas, juntamente com uma explicação do processo de captura. È ainda efetuada uma descrição das TF e TW, bem como uma discussão sobre as vantagens e limitações associadas a cada abordagem de decomposição. O capítulo prossegue com algumas definições importantes que serão utilizados intensivamente nos próximos capítulos. Por fim, é apresentado o cenário de medição, os meios utilizados nos testes e as condições em que estes foram efetuados. É ainda referido como foram obtidos os dados para os vários casos estudados, apresentando os programas utilizados e as aplicações *Web* analisadas.

3.1 Aplicações Internet, Tráfego Internet

Internet

A Internet é uma rede global de redes interligadas, constituída por biliões de utilizadores no mundo inteiro. Como já foi mencionado no capítulo 1, nos últimos anos esta rede tem crescido em tamanho, complexidade e importância. Além disso, o aumento recente e impressionante dos serviços e aplicações implica o desenvolvimento de novos paradigmas de comunicação e transporte dos diferentes tipos de dados, tais como ficheiros, voz, vídeo, entre muitos outros. Segundo as previsões da Cisco (Figura 3.1), o tráfego IP global deverá quadruplicar entre 2011 e 2016, crescendo até 1.3 Zettabytes por ano (110 Exabytes por mês) em 2016, sendo que cerca de 87% deste tráfego será de consumo residencial. Em 2016, dos 110 Exabytes do tráfego da Internet gerados por mês, cerca de metade deverão ser originados por aplicações de vídeo.

O tão conhecido protocolo de Internet (IP) é utilizado para associar e transportar os dados entre os computadores ligados e pode ser visto como uma linguagem universal da Internet, compreendida por vários computadores e dispositivos, permitindo assim a interligação de vá-

rias redes. Desde o final do século XX que as descobertas tecnológicas se têm intensificado a uma velocidade tal que alteraram as relações entre as sociedades. As tecnologias da Informática e das Telecomunicações convergiram numa única rede de comunicação: a Internet. Com a expansão da Internet as distâncias foram reduzidas. Essa dinâmica facilitou o acesso a informações e produtos, forçando uma reestruturação económica e a necessidade de adaptação à nova indústria de conteúdos acessíveis ao clicar de um rato. Muitas empresas viram-se obrigadas a divulgar os seus produtos e serviços na Internet através de um *website*.

Apesar de o protocolo IP ser capaz de transmitir todos os pacotes, não garante a sua entrega eficaz e atempada aos destinatários. Estas tarefas têm que ser asseguradas por outros protocolos específicos da camada de transporte, como o TCP, que garante uma entrega fiável e ordenada dos pacotes solicitados. O encapsulamento das diferentes camadas de rede e dos seus protocolos é um conceito muito importante na rede, o que leva à criação de diferentes componentes de frequência no tráfego da Internet [54].

A evolução contínua da Internet é muito mais rápida do que a nossa capacidade para a caraterizar, entender, controlar ou prever. A área de pesquisa da classificação de tráfego da Internet inclui muitos trabalhos que representam as várias tentativas de caraterizar e classificar o tráfego.

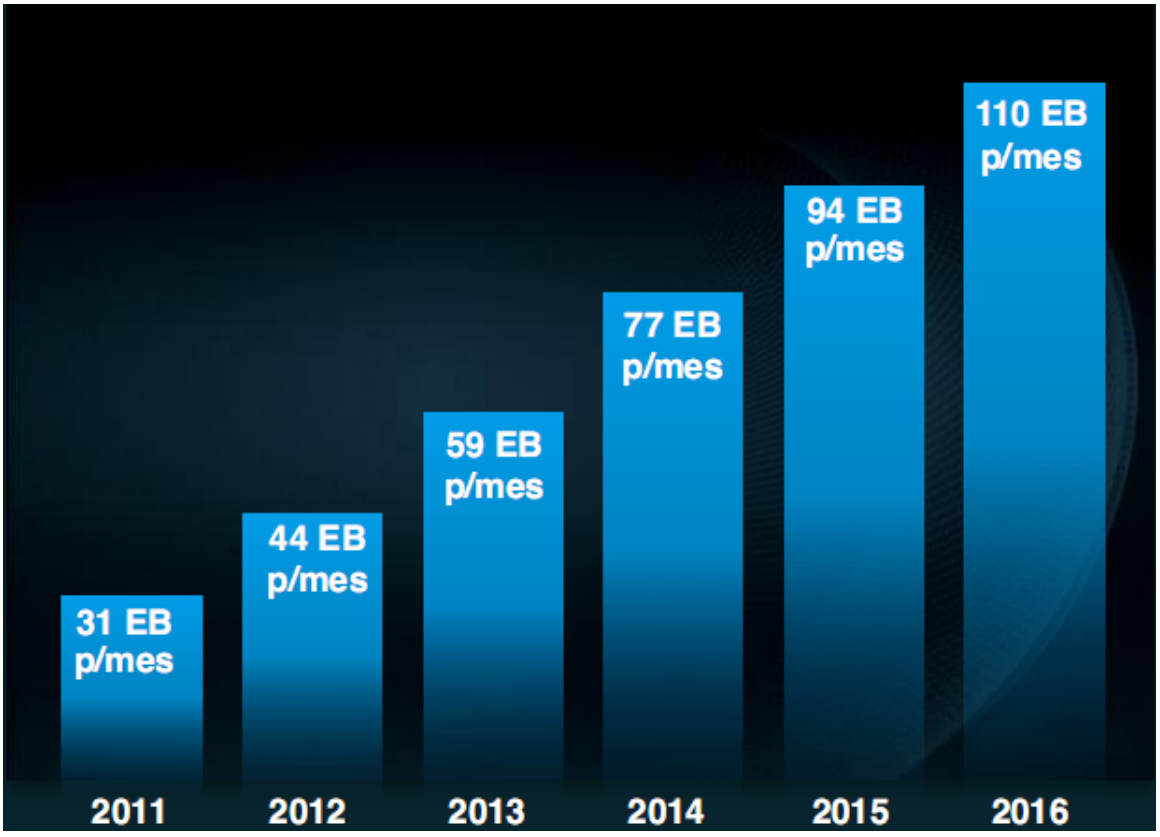


Figura 3.1: Crescimento do Tráfego IP, 2011-2016 [retirado de Cisco VNI Global Forecast *[2]*].

Aplicações Internet

O desenvolvimento de aplicações para a Internet tem sido uma das maiores tendências no ramo de desenvolvimento de software. Nos últimos anos tem crescido com uma velocidade espantosa, abrangendo todas ou quase todas as áreas de negócios imagináveis. A figura 3.2 mostra a previsão de crescimento das aplicações da Internet para os anos 2011-2016, podendo-se observar que as aplicações de Vídeo dominam o tráfego dos consumidores. As aplicações de Internet têm contribuído muito para o desenvolvimento de diversas áreas da Ciência da Computação, tendo ainda aumentado a capacidade, potencial, qualidade e diversidade dos serviços que são oferecidos no mercado atual e podem vir a ser oferecidos no futuro.

Um aspeto muito importante destas aplicações é que cada aplicação requer interação de diferentes utilizadores, de acordo com o serviço utilizado, e gera diferentes interações com *hosts* e servidores remotos que levam à criação de dinâmicas de tráfego diferentes.

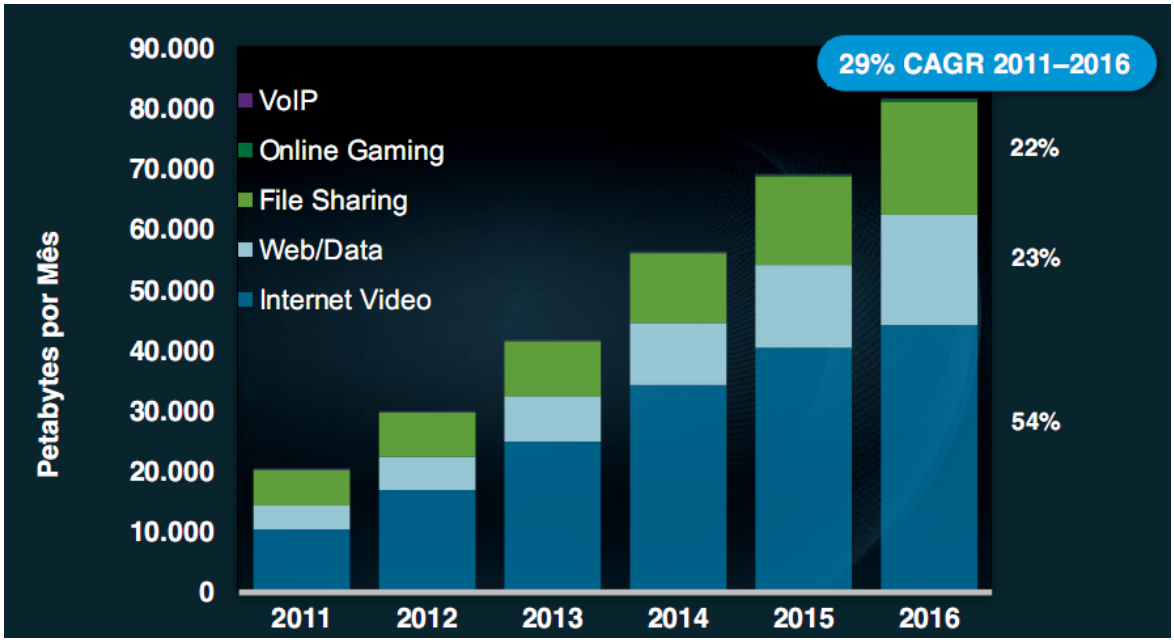


Figura 3.2: Crescimento das aplicações da Internet, 2011-2016 [retirado de Cisco VNI Global Forecast *[2]*].

Tráfego Internet

A interação dos utilizadores com as aplicações *web* podem originar picos de tráfego diferentes, correspondentes a solicitações de páginas *web*, *download* de ficheiros, etc. O consumo de largura de banda varia conforme a aplicação requerida pelo utilizador: por exemplo, as aplicações de vídeo, por não requererem interações tão frequentes do utilizador, apresentam um consumo de largura de banda constante e considerável, devido à transmissão do vídeo. Estas aplicações são sensíveis a atrasos e *jitter*, que afetam a qualidade de receção e perceção do vídeo. Por outro lado, os vários eventos e mecanismos de rede constroem as diferentes componentes de frequência do tráfego Internet. Um utilizador da Internet ao solicitar uma

página *web*, cria um conjunto de sessões que, por sua vez, criam um conjunto de pacotes que são transmitidos através da ligação física. Estes eventos criam várias componentes de frequência em diferentes regiões do espetro. Este conceito é ilustrado na figura 3.3, que mostra três regiões diferentes do espetro de frequência, juntamente com seus eventos. Componentes de frequência baixa constituem eventos humanos, que no mundo da Internet estão associados aos comportamentos e ações humano/utilizador, como cliques de um utilizador numa página *web*. Entre as regiões de baixa e alta frequência existe uma região de frequência de gama média que corresponde a eventos de rede, tais como criação de sessões de tráfego e respetivos mecanismos de controlo. Outros mecanismos de controlo utilizados, tais como modelação de tráfego, são também abrangidos por esta região. Finalmente, na região do espetro correspondente ás altas frequências são tidos em conta eventos dos protocolos e da Internet, como chegadas de pacotes. Aplicações de Internet que apresentam estas componentes são as que geram uma quantidade considerável de tráfego com um elevado número de pacotes recebidos.

Todos estas componentes de frequência estão associadas ás diferentes interações simultâneas que são geradas pelas aplicações de Internet que correm nos vários clientes e servidores remotos. A análise destas componentes é crítica para conseguir uma diferenciação eficiente entre as dinâmicas geradas pelas diferentes aplicações. Na sub-secção seguinte será apresentada uma definição importante para a resolução deste problema.

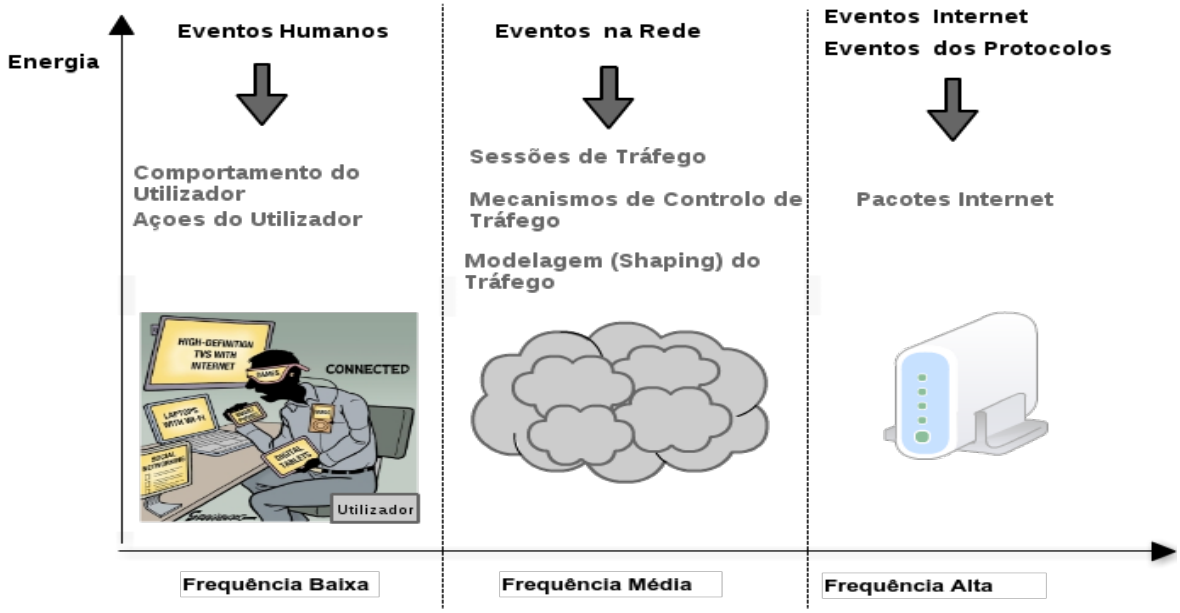


Figura 3.3: Mapeamento entre as diferentes regiões do espetro de frequência e os eventos de rede e dos utilizadores [editado de [3]].

3.1.1 Data-Streams

Como foi mencionado anteriormente, uma aplicação Internet gera várias interações simultâneas com *hosts* e servidores remotos que conduzem à criação de diferentes dinâmicas de tráfego.

Tecnicamente, um *Data-Stream* pode ser visto como um conjunto contínuo de fluxos de

dados que são transmitidos ou recebidos de um *host* local. Note-se que, tradicionalmente, o tráfego Internet é agrupado em fluxos de acordo com a definição clássica *five-tuple* (endereço IP origem e destino, números das portas de origem e destino, e protocolo da camada de transporte).

No entanto, estes fluxos correspondem apenas a uma das muitas interações geradas pelas aplicações da Internet. No sentido de ser capaz de analisar e classificar tais interações e as suas diferentes componentes de frequência, a definição clássica e restritiva de fluxo será substituída pela definição de *data-stream*. Um data-stream consiste em todo o tráfego (nas direções *upload* ou *download*) de um endereço IP local, que é univocamente identificado por um identificador numérico. Este identificador numérico pode ser um número de porta TCP / UDP específica, para tráfego não cifrado, enquanto que para tráfego cifrado pode ser o índice do parâmetro de segurança do cabeçalho *ESP*, no caso de túneis *Internet Protocol Security* (IPSec), ou qualquer outro identificador específico de tecnologia de túneis cifrados do nível IP.

Os *data-streams* são assim identificados por um *2-tuple* (Endereço IP, identificador único). A análise das componentes de frequência pode desempenhar um papel importante na discriminação do tráfego. O conceito de *data-stream* é ilustrado na figura 3.4, que mostra uma comparação entre *data-streams* e fluxos da Internet.

Os *data-streams* que serão analisados nesta dissertação foram recolhidas na Universidade de Aveiro, tendo sido utilizado o método DPI para efetuar a sua validação. A definição de *data-stream* é utilizada uma vez que a nossa intenção é classificar as diferentes interações do utilizador da Internet, que podem consistir em várias ligações simultâneas com diferentes *hosts* ou servidores. Os *data-streams* recolhidos correspondem a *downloads* com duração de 5 minutos, amostrados em intervalos de 0.1 segundos (3000 amostras).

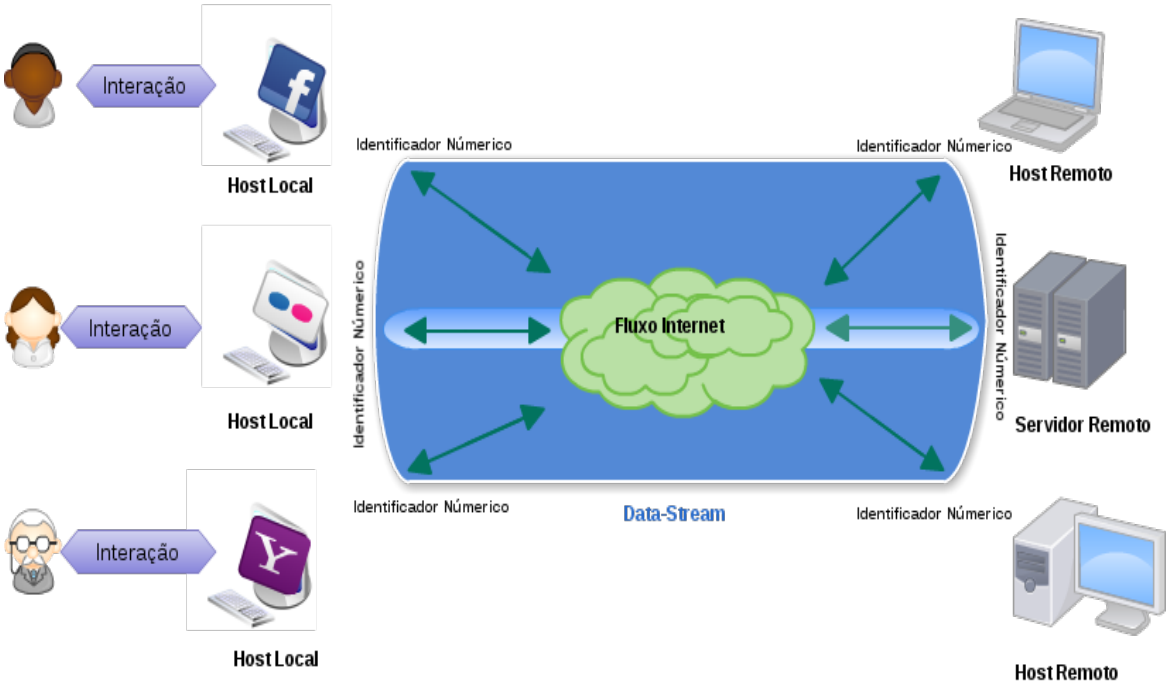


Figura 3.4: Tráfego gerado por uma aplicação da Internet: *data-streams* vs fluxos. [editado de [3]].

3.1.2 Traces de Tráfego

Esta sub-seção apresenta os *traces* de tráfego reais que foram capturados para avaliar a precisão das diferentes metodologias propostas. O tráfego das várias aplicações de Internet estudadas foi passivamente recolhido na rede de comunicações da Universidade de Aveiro, tendo sido medido entre Outubro de 2011 e Setembro de 2012. O tráfego capturado é composto por pacotes IP / TCP e IP / UDP cifrados, tendo sido utilizado o Wireshark, publicamente disponível em [55], para capturar o cabeçalho completo e os primeiros 64 bytes do *payload* de cada pacote. A fim de verificar a capacidade dos diferentes métodos de classificação propostos e das metodologias de identificação, o tráfego foi dividido em duas categorias:

1. *Traces* de treino: usados no estudo do tráfego original da aplicação Internet;

2. *Traces* de teste: usados para avaliar a capacidade de identificar e classificar o tráfego legítimo da aplicação Internet.

Na sub-seção seguinte, será apresentado o tráfego das aplicações legítimas que serão estudadas, como foram efetuadas as capturas, quais os protocolos que foram utilizados e quais os padrões de tráfego que foram obtidos.

3.1.3 Serviços

De modo a validar a capacidade das metodologias de classificação propostas para a identificação precisa das aplicações, foram utilizadas cinco classes de serviços *on-line*:

- **redes sociais:** o tráfego de redes sociais foi gerado usando uma conta criada num *website*, interagindo com as atualizações de notícias provenientes dos demais utilizadores ligados, o que incluía comentários e "likes";

- **notícias *on-line*:** o tráfego de notícias *on-line* foi gerado através da visita ao site de um jornal;

- **e-mail on-line:** o tráfego *e-mail on-line* foi gerado usando os serviços oferecidos por um serviço de *e-mail*, concretamente o tráfego gerado apenas pelas sincronizações automáticas entre o terminal *web* do cliente e o servidor;

- **Partilha de fotos:** para gerar tráfego de uma aplicação de partilha de fotos *on-line*, foi criada uma conta num *website* de partilha de fotos, tendo apenas sido considerado para análise o tráfego gerado enquanto se navegava entre as fotos de outros utilizadores;

- **serviços de vídeos:** o tráfego de *download* de vídeo *on-line* foi gerado assistindo a vídeos num *website* de visualização e partilha de vídeos.

Na seção seguinte, serão apresentadas as classes e os comportamentos caraterísticos dos diversos serviços da Internet.

Serviços de Redes Sociais

As redes sociais, para além de reunir pessoas amigas e colegas de trabalho, estudo e convívio, fornecem a possibilidade de cada pessoa poder ter o seu perfil, ou seja, os seus dados pessoais, tais como fotos, vídeos, *links*, notas, etc. Atualmente as redes sociais fazem parte do dia-a-dia das pessoas e das empresas em todo o planeta. Desde o aparecimento do Orkut e com o crescimento do Facebook e do Twitter e do recentemente criado Google+, que permitem acessos por computadores fixos e por dispositivos móveis, que as redes sociais se transformaram numa autêntica febre mundial. Muitas empresas, devido ao crescimento exponencial de utilizadores das redes sociais, passaram a expor as suas marcas e produtos através delas como uma estratégia de marketing para os tornar mais acessíveis aos clientes. Trata-se de uma forma interativa e em tempo real de estar em constante contato com o cliente, que pode ainda apresentar as suas sugestões e reclamações. Este estreitamento da relação entre público e marca ajuda o empresário a identificar questões importantes do seu negócio. Através do *feedback* do público, a empresa consegue traçar perfis e definir novas estratégias. O numero de utilizadores de rede sociais ascende aos 301 milhões.

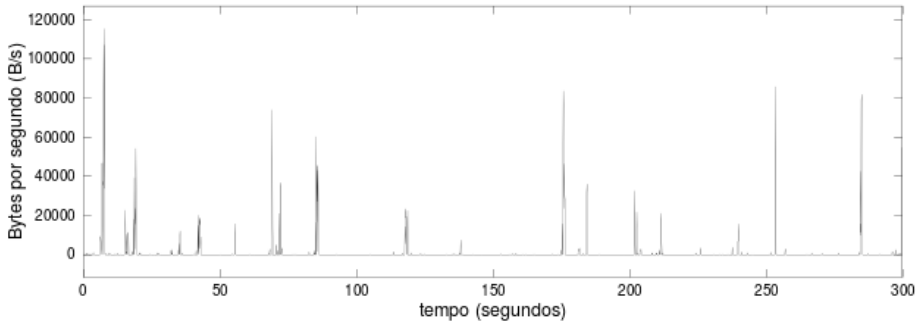
O tráfego gerado pelas redes sociais, que se pode observar na figura 3.5, apresenta picos frequentes, que são gerados pelas atualizações do estado criadas por outros utilizadores ligados, que normalmente consistem apenas em mensagens, e também pela solicitação de uma nova página (perfil), comentários e "likes". Os intervalos de tempo que não apresentam tráfego correspondem ao período em que o utilizador não efetua nenhuma ação, períodos em que o utilizador pode estar apenas a visualizar uma fotografia ou a ler uma determinada notificação.

Facebook

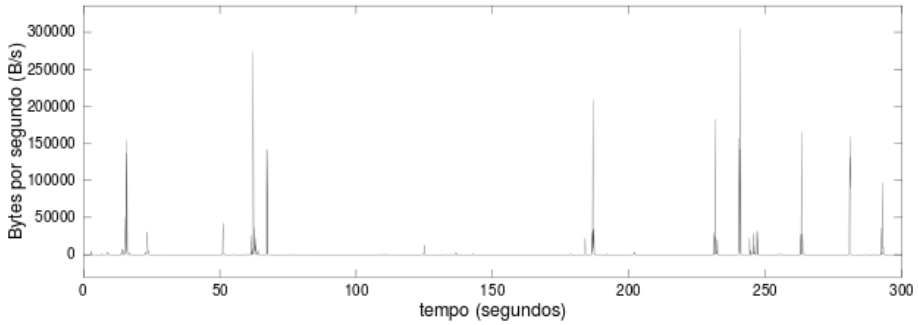
Neste *site* cada pessoa pode ter o seu perfil, ou seja, os seus dados pessoais, as suas fotos, vídeos, *links*, notas, etc. A interação entre os membros desta rede social pode ser feita visitando os perfis, fazendo amigos, estabelecendo contatos, deixando comentários, enviando mensagens, numa palavra, comunicando. O Facebook foi fundado por Mark Zuckerberg em 2004, inicialmente com o objetivo de ajudar os estudantes da Universidade de Harvard, mas progressivamente foi permitindo o registo de utilizadores de outras escolas, até que em 2006 se tornou disponível para todos. Para usar o serviço é necessário ter uma conta no Hotmail, com um perfil registado. Atualmente o Facebook conta com mais de 800 milhões de utilizadores.

Google$^+$

O Google$^+$ é um projeto recente, criado pela Google, com o principal objetivo de melhorar a privacidade do utilizador nas redes sociais, através da construção de grupos específicos de amigos, chamados "Círculos". Esta rede social é considerada um mistério, mas tem tido uma adesão substancial ao longo dos anos. Permite, entre várias outras coisas, uma videoconferência com várias pessoas ao mesmo tempo, partilha de conteúdos, organização de interesses por círculos e, o melhor de tudo, é que as páginas desta rede social aparecem nas pesquisas do Google. Para usar o serviço é necessário ter uma conta no Google, com um perfil registado. Por ser um serviço recente, o numero de utilizadores do Google$^+$ está entre os 170 e os 180 milhões.



(a) Tráfego Facebook - direção download.



(b) Tráfego Google^{+} - direção download.

Figura 3.5: Tráfego das Redes Sociais *On-line* Facebook 3.5a e Google^{+} 3.5b.

Serviços de notícias *on-line*

Falar da evolução da Internet e das tecnologias implica obviamente falar das transformações dos meios de comunicação, nomeadamente do Jornalismo *on-line*. A Internet é, neste momento, o meio de comunicação que mais torna visível a convergência dos media num único suporte. A migração dos meios jornalísticos impressos para a Internet trouxe muitas vantagens aos leitores, fornecendo além do acesso aos conteúdos do dia a possibilidade do leitor interagir com os conteúdos através de motores de busca e da navegação.

O tráfego dos serviços de notícias *on-line* (figura 3.6) apresenta vários picos aperiódicos de curta duração. Esses picos são causados pelo utilizador quando clica em *hiperlinks* enquanto navega através das notícias disponíveis, fazendo *download* de uma nova página que apresenta a notícia solicitada. Os intervalos de tempo onde não há tráfego correspondem ao período em que o utilizador permanece numa determinada página a ler uma notícia específica.

Abola

Abola (www.abola.com) é um diário português de informação generalista especializado em desporto, que ocupa a décima primeira classificação nos *Top Sites* de Portugal. Para além de edições impressas do jornal, Abola faz uso das edições *on-line*, dando ao utilizador a possibilidade de interagir com o jornal e até de notificar e funcionar como fonte de informação; oferece ao utilizador a possibilidade de navegar por outros *sites* através de *hiperlinks*; faz uso de hipermédia, ou seja, a união num único suporte de conteúdos escritos, sonoros e

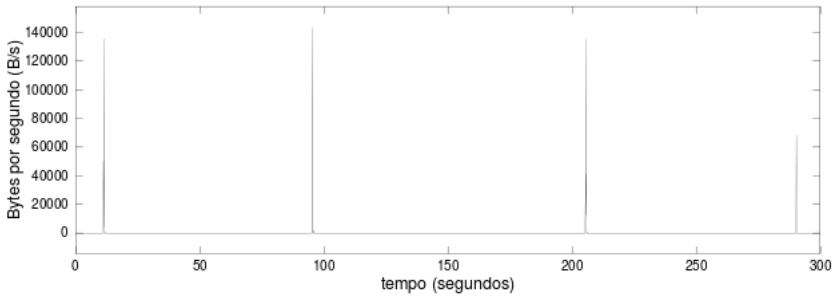
imagéticos, seja as imagens fixas ou animadas; permite ainda que o utilizador tenha as notícias instantaneamente.

Record

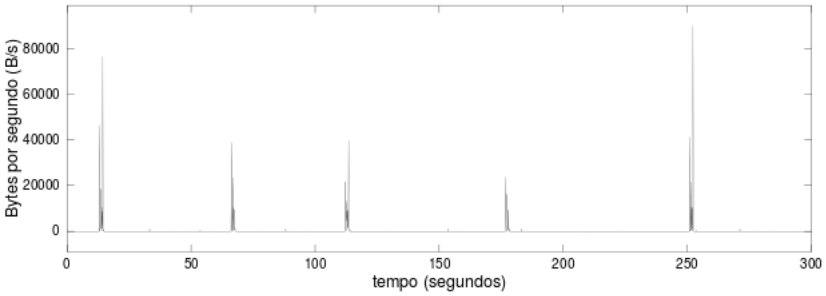
O Record (www.record.pt) é um diário português de informação desportiva e apresenta muitos aspetos semelhantes ao serviço de notícias *on-line* apresentado anteriormente (Abola).

CNN

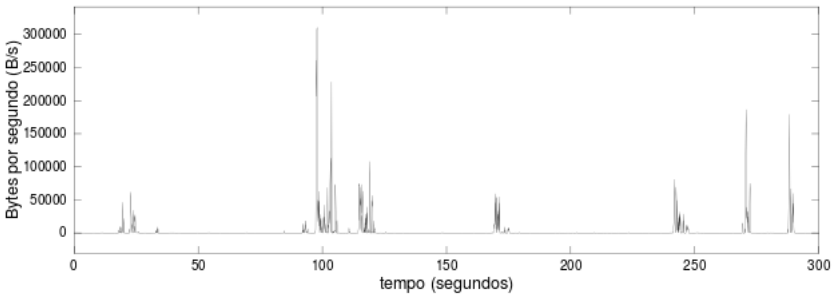
A CNN (www.cnn.com) é um jornal de notícias *on-line* dos Estados Unidos, fundado em Agosto de 1995, que fornece notícias de todo o mundo e informações de negócios, desportos, etc. É um dos sites de notícias mais populares do mundo.



(a) Tráfego de Abola - direção download.



(b) Tráfego de Record - direção download.



(c) Tráfego da CNN - direção download.

Figura 3.6: Tráfego dos Serviços de Notícias *On-line* Abola 3.6a, Record 3.6b e CNN 3.6c.

Serviços de e-mails

O serviço de *e-mail* tornou-se o principal meio de comunicação para a maioria das organizações, sendo uma ferramenta de comunicação essencial para todos os utilizadores. O *e-mail* é muito utilizado como repositório de informações comerciais e muitas organizações utilizam este serviço para realizar transações comerciais, como envio de contratos ou ordens de compra. Os serviços de *e-mail* são utilizados por mais de 279 milhões de pessoas.

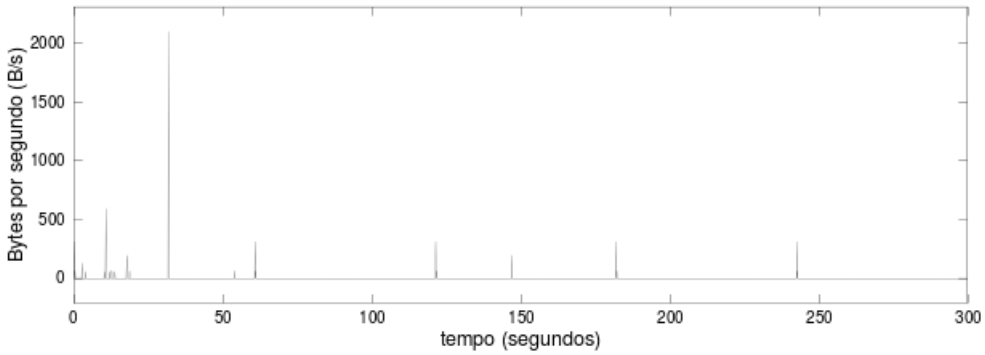
O tráfego gerado pelos serviços de *e-mail*, como mostra a figura 3.7, apresenta poucos picos e pouco frequentes, que correspondem a sincronizações inicial e automáticas entre o terminal *web* do cliente e o servidor do Hotmail. Estes picos têm duração muito curta e são pouco frequentes, porque o tráfego de sincronização apenas verifica a existência de novos *e-mails*.

Hotmail

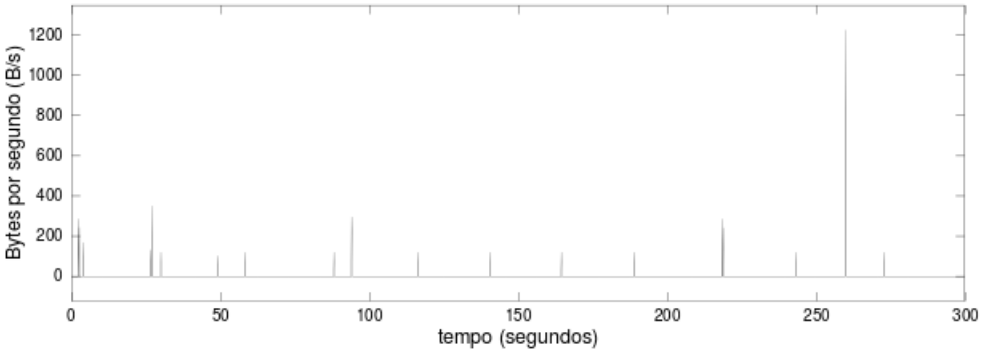
O Hotmail (www.hotmail.com) é um serviço de correio eletrónico (*e-mail*) gratuito da empresa Microsoft. O Windows Live Hotmail, como é oficialmente conhecido o Hotmail, é o nome do serviço de correio eletrónico gratuito que é acedido através da *web*. Foi um dos primeiros serviços de *webmail* (*e-mail* via *web*). Criado pelo indiano Sabeer Bhatia com ajuda de Jack Smith, foi lançado em 4 de julho de 1996, tendo sido posteriormente comprado pela Microsoft. Atualmente o serviço conta com mais de 360 milhões de utilizadores em todo o mundo, e é considerado um dos maiores serviços de *e-mail*.

Gmail

O Gmail (www.gmail.com) é um serviço de *webmail* gratuito baseado em pesquisa, que combina os melhores recursos de *e-mail* tradicionais com a tecnologia de pesquisa do Google. O Gmail, além de oferecer uma forma completamente nova de ler e rastrear mensagens, oferece mais espaço para armazenamento. Atualmente a Google anunciou que o seu popular serviço de *e-mail* atingiu mais de 425 milhões de utilizadores ativos espalhados pelo mundo, tornando-se no maior serviço de *e-mail* do mundo, ultrapassando o seu eterno concorrente Hotmail.



(a) Tráfego Hotmail - direção download.



(b) Tráfego Gmail - direção download.

Figura 3.7: Tráfego dos Serviços de *E-mail On-line* Hotmail 3.7a e Gmail 3.7b.

Serviços de Partilha de Fotos

Os serviços de partilha de fotos *on-line* estão a ganhar mais utilizadores. Para além de possibilitarem a criação de álbuns fotográficos *on-line*, também oferecem a possibilidade de criar "slide shows"com bom aspeto sem recorrer ao PowerPoint.

As aplicações de partilha de fotos geralmente geram vários picos de tráfego com pseudo-periodicidade, como se pode observar na figura 3.8, devido aos cliques que são executados pelo utilizador enquanto pede para ver outra fotografia. Os intervalos de tempo onde não há tráfego correspondem ao período que o utilizador permanece numa determinada página a visualizar uma fotografia.

Flickr

O Flickr (www.flickr.com) é um *site* da Internet que permite armazenar e partilhar fotografias, entre outros documentos gráficos, como desenhos e ilustrações, além de permitir várias formas de organizar fotos e vídeos. É caraterizado como uma rede social, oferecendo aos seus utilizadores a possibilidade de criarem álbuns para armazenamento das suas fotografias e entrarem em contacto com fotógrafos variados e de diferentes locais do mundo.

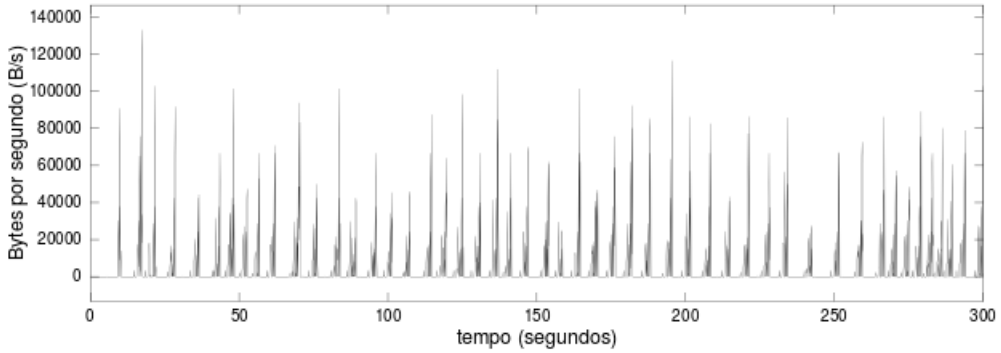
O *site* foi criado pela Ludicorp em Vancouver, Canadá, sendo lançado em fevereiro de 2004. No começo de 2005, o *site* foi adquirido pela Yahoo. Atualmente o Flickr tem mais de

32 milhões de utilizadores. O Flickr é provavelmente o melhor aplicativo de gestão e partilha de foto e vídeos *on-line* e fornece novas formas de organização de fotos e vídeos.

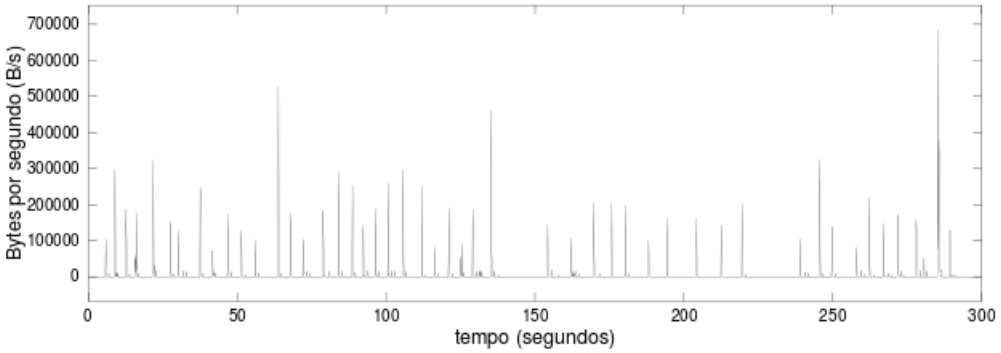
Picasa

Semelhante ao Flickr, o Picasa (picasaweb.google.com) permite o armazenamento e gestão de fotos/imagens *on-line*, tendo as vantagens de pertencer à Google (interacção facilitada) e de ter utilitários que permitem a edição dos conteúdos. Lançado pela Google, o Picasa Web Album é concebido para que os utilizadores possam colocar as suas fotos na Internet de modo a que sejam partilhados com outros utilizadores que tenham uma conta no *webmail* do Google, o Gmail. Contudo, para visualizar as fotos não é necessário ser utilizador do Gmail.

O Picasa Web Album permite ainda que todos os amigos e contatos sejam notificados sobre os novos álbuns disponíveis através de um *link* especial que os utilizadores podem enviar a quem quiserem. Esta nova funcionalidade faz com que o Google possa concorrer com o popular Flickr, um serviço de partilha *on-line* de fotos ligado ao portal concorrente Yahoo.



(a) Tráfego Flickr - direção download.



(b) Tráfego Picasa - direção download.

Figura 3.8: Tráfego dos Serviços de Partilha de Fotos *On-line* Flickr 3.8a e Picasa 3.8b.

Serviços de Partilha de Vídeos

O conceito de partilha de Vídeo (vídeo sharing) já existe à cerca de 5 anos, permite aos utilizadores a visualização, armazenamento e partilha dos seus vídeos recorrendo a um *website*, que posteriormente disponibiliza os vídeos através do seu servidor para outros utilizadores.

O aparecimento de *sites* de partilha de vídeos deu-se em meados de 2005, e ao longo dos anos tem evoluído de uma forma exponencial, contando atualmente com um número significativo de *sites*. Os serviços de partilha de vídeos acompanharam a crescente evolução tecnológica, existindo atualmente vários tipos de formato e codificação dos vídeos que melhoram a qualidade de imagem. Estes serviços fornecem várias funcionalidades ao utilizador, como a possibilidade de criar *playlists*, comentar vídeos, votar nos vídeos e muito mais. Nos últimos anos a importância dos serviços de partilha de vídeo aumentou. Hoje em dia, em cada segundo que passa são carregadas 24h de vídeo nos *websites* de partilha de vídeos. Muitas empresas usam esta estratégia para promover o seu serviço ou produto, demonstrando ser este um meio eficiente de *markting*. A partilha de vídeos tem crescido a um ritmo elevado nos últimos anos: desde de 2006, a partilha de vídeos *on-line* mais do que triplicou, e é considerada a atividade *on-line* mais frequente no mundo. Pode-se dizer que atualmente a maioria das pessoas prefere ver vídeos *on-line* do que ler um texto extenso, o que mostra a importância destes serviços.

Pode-se considerar dois tipos de vídeos:

- **Vídeos de Longa Duração (VLD) (figura 3.9):** estes vídeos têm uma duração maior, incluindo filmes, séries, jogos completos de futebol,etc.
- **Vídeos de Curta Duração (VCD) (figura 3.10):** são vídeos com uma duração pequena, como por exemplo vídeos de musica, resumos de jogos de futebol, etc.

As figuras 3.9 e 3.10 ilustram os padrões de comportamento relacionados com capturas de vídeo de longa e curta duração.

O comportamento do tráfego depende das caraterísticas dos *sites* e da importância que estes dão a cada vídeo. Através das figuras pode-se observar que nem sempre os vídeos apresentam o mesmo comportamento, sendo que apesar de serem serviços de partilha de vídeos, o comportamento dos VLD e de VCD é diferente.

Por exemplo, o modo de funcionamento do Youtube em relação à transferência dos vídeos difere de muitos *sites* do mesmo género. O Youtube altera o tamanho do *buffer* dos vídeos conforme a sua duração, a sua popularidade e a qualidade com que são transmitidos.

Os vídeos com mais popularidade requerem mais atenção na sua correta transmissão e para isso são necessários mais recursos, são colocados em vários servidores por forma a servirem o utilizador com uma melhor qualidade de serviço.

Nos vídeos com qualidade bastante alta (HD) o tráfego transferido é muito elevado e o carregamento é feito com maior velocidade, diminuindo substancialmente o tempo de carregamento em relação ao que acontece com os vídeos normais.

É possível verificar a existência de duas zonas distintas na figura 3.10, correspondendo a dois vídeos diferentes. Os intervalos de tempo onde não há tráfego advém do período entre a escolha e alteração (normal para HD) de um vídeo.

O tráfego de *download* dos vídeos *on-line* foi gerado assistindo a vídeos em HD. Estes serviços geram tráfego com alguns picos e com alguma frequência, como mostra as figuras 3.9 e 3.10, uma vez que o número de cliques do utilizador não é tão relevante.

Youtube

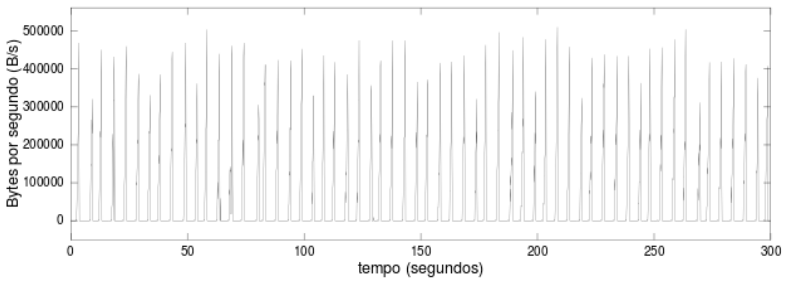
O YouTube (www.youtube.com) criado por Chaud Hurley é um *site* de partilha de vídeos que ganhou bastante notoriedade, tornando-se um fenómeno da Internet. É atualmente o *site* de partilha de vídeos com maior fluxo de utilizadores e com o maior número de vídeos partilhados. Fundado em fevereiro de 2005, tem como principais funcionalidades o facto dos utilizadores registados poderem interagir com o sistema, proporcionando ao utilizador a possibilidade de criar canais de TV com os seus vídeos, poder ter vídeos favoritos de outros utilizadores e até listas de reprodução automática. Recentemente tornou-se possível a reprodução em vídeos com qualidade HD num formato de 1280x720; os *uploads* podem ser feitos em vários formatos até 15 minutos e com o máximo de 2 Gb de tamanho. Hoje em dia o YouTube disponibiliza séries e filmes completos e gratuitos para o público.

Nas figuras 3.9a e 3.10a é possível analisar a forma como o tráfego de vídeos com qualidade bastante elevada é recebido, podendo-se observar que nos VLD o tráfego transferido é muito elevado e constante ao longo do tempo, ao contrário do que acontece nos VCD, em que apesar do tráfego transferido ser muito elevado é bastante irregular.

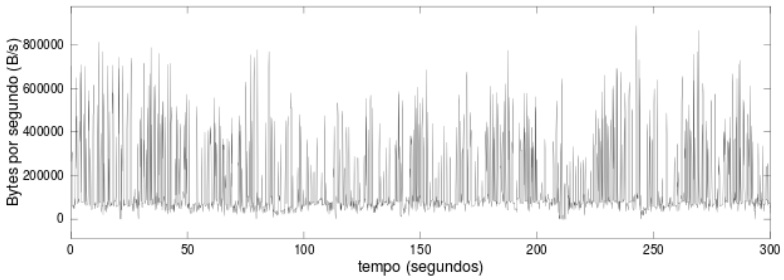
Vimeo

O Vimeo (www.vimeo.com) é outro serviço popular de visualização e partilha de vídeos. Difere de outros serviços por não permitir vídeos comerciais, sendo muito utilizado pelas comunidades mais artísticas. Foi criado por Jake Londwick e Zack Klein em novembro de 2004, sendo atualmente um dos sites de partilha de vídeos com mais utilizadores (mais de sete biliões). É de livre acesso e permite fazer *upload* até 500 MB por semana, também tem a opção paga que permite o *upload* de 5 GB por semana para além de muitas ofertas como *upload* ilimitado de vídeos em HD. Ao contrário de outros sites de partilha de vídeos, o Vimeo permite fazer *upload* de vídeos de duração bastante longa e em formato de alta definição, o que atrai um número significativo de especialistas em animação, cinema, etc.

Analisando as figuras 3.9b e 3.10b, pode-se observar o comportamento dos vídeos do Vimeo, chegando-se à conclusão que o Vimeo opera a um ritmo bastante irregular. A média da velocidade de transmissão difere conforme a qualidade (HD ou normal) e duração dos vídeos.

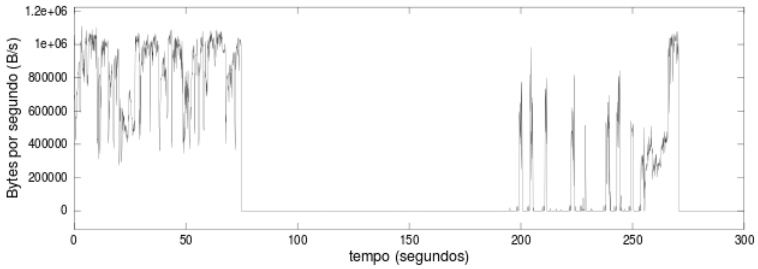


(a) Tráfego Youtube - direção download.

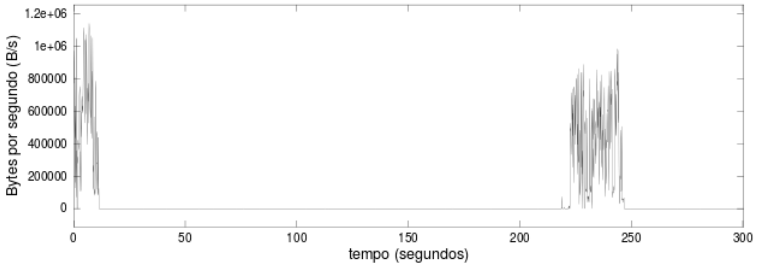


(b) Tráfego vimeo direção - download.

Figura 3.9: Tráfego dos Serviços de Partilha de Vídeos HD *On-line* (VLD) Youtube 3.9a e vimeo 3.9b.



(a) Tráfego Youtube - direção download.



(b) Tráfego vimeo - direção download.

Figura 3.10: Tráfego dos Serviços de Partilha de Vídeos HD *On-line* (VCD) Youtube 3.10a e vimeo 3.10b.

3.2 Análise escalar do tráfego

As ferramentas de análise de qualquer processo nos domínios do tempo e da frequência são amplamente utilizadas em muitas áreas diferentes, como a análise de imagem, a compressão de dados, e mais recentemente a análise de tráfego. Nesta seção, são discutidas as ferramentas mais relevantes para a análise dos componentes de frequência de sinais e séries temporais.

Começamos por apresentar a TF, a técnica mais utilizada na análise de espetros de frequência de processos estocásticos, descrevendo a sua fórmula matemática. São também apresentadas as vantagens e desvantagens associadas a esta abordagem. Posteriormente, será feita uma descrição da TW, um método de análise de sinal que permite resolver alguns dos problemas associados ás TFs. Finalmente, apresentaremos a nossa abordagem multi-escalar de análise de tráfego, explicando como as *wavelets* são utilizadas nesta abordagem.

3.2.1 Transformada de Fourier

A TF é a técnica mais utilizada para efetuar a análise do espetro de frequência de processos estocásticos, decompondo-o em funções exponenciais complexas com frequências diferentes [56].

Definamos $L^2(R)$ como o conjunto de funções quadráticas e integráveis, *i.e*, o conjunto de funções reis *x(t)* que satisfazem:

$$\int_{-\infty}^{+\infty} |x^2(t)|dt < \infty \tag{3.1}$$

em que o produto interno é definido como:

$$\langle x, y\rangle = \int_{-\infty}^{+\infty} x(t)y^*(t)dt \tag{3.2}$$

com norma:

$$\parallel x \parallel = \langle x, y\rangle^{1/2} \tag{3.3}$$

A TF da função $x(t) \in L^2(\mathbb{R})$ pode ser definida como:

$$\mathbf{X}(w) = \int_{-\infty}^{+\infty} x(t)e^{-iwt} \tag{3.4}$$

onde w representa a frequência da sinusóide analisada. Uma vez que o suporte da sinusóide não é localizad, a TF apresenta uma resolução temporal pobre, sendo apenas adequada para a análise de sinais estacionários, ou seja, sinais que apresentam a mesma componente de frequência em toda a gama de análise. Consequentemente, as TFs são incapazes de fornecer uma representação tempo-frequência onde os diferentes componentes de frequência de um processo não estacionário sejam descritos juntamente com os intervalos de tempo em que ocorrem. Portanto, a análise de sinais variáveis no tempo com alterações bruscas exige outras ferramentas de análise [56].

3.2.2 Wavelets

Tal como mencionado, a TF requer que o sinal analisado seja estacionário, isto é, que as componentes de frequência dos dados analisados não mudem ao longo do tempo. Em muitos casos, essa restrição é respeitada pelos dados e pode ser obtida uma decomposição precisa. No entanto, este não é o caso do tráfego da Internet, que é conhecido por ser não estacionário já que apresenta componentes de frequência diferentes em intervalos de tempo diferentes.

As TWs, pelo facto de admitirem que o sinal analisado possa ser não estacionário [57], são capazes de fornecer uma representação tempo-frequência do sinal e são amplamente aplicadas em diversas áreas, tais como processamento de sinal, análise e compressão de imagem. Trata-se de uma técnica poderosa para a compreensão da complexidade de diversos processos do mundo real.

Wavelets são funções matemáticas que são utilizadas para dividir um sinal em diferentes componentes de frequência. Foram inicialmente apresentadas em 1980 pelo geofísico J.Morlet com o objetivo de efetuar decomposição e aproximação de sinais, e consistem num curto período de oscilação semelhante a uma onda com uma amplitude limitada, que ocorre num curto período de tempo, o que lhe confere a capacidade de obter uma boa resolução em termos de tempo e frequência. A TW também pode ser vista como um processo de convolução entre os dados analisados e uma versão alterada e estendida de uma wavelet. A operação de translação que é aplicada sobre a wavelet permite deslocá-la através do sinal, enquanto o fator de escala permite esticar ou comprimir a wavelet, dando a possibilidade de analisar as diferentes componentes de frequência do sinal. Em escalas menores, a TW atua como um *zoom in* (aumento da imagem) sobre a função, revelando detalhes de alta frequência nos dados analisados. Por outro lado, em escalas maiores a TW atua como um zoom out (diminuição da imagem) na função, que é adequado para a análise dos componentes de baixa frequência dos dados.

As wavelets permitem a análise de cada um dos componentes do sinal numa escala apropriada e apresentam várias vantagens sobre as outras técnicas de análise de sinal, como as TFs. As TFs são mais adequados para analisar dados periódicos, enquanto que as TWs são mais apropriadas para análise de funções com descontinuidades e picos. Como as wavelets apresentam um suporte compacto, permitem uma resolução de tempo muito boa e podem, consequentemente, fornecer informações sobre o tempo e a frequência, enquanto que as TFs apenas fornecem informações da frequência. Além disso, o conjunto infinito de funções básicas das wavelets é outra das vantagens que apresentam sobre a TF, que utiliza um conjunto finito de funções básicas (senos e cossenos).

Segundo [[58], [59]] uma wavelet $\psi(t)$ pode ser definida como uma função passa-banda, oscilando numa frequência central f_0, que satisfaz a seguinte condição de admissibilidade:

$$0 < C_\Psi = 2\pi \int_{-\infty}^{+\infty} \frac{|\Psi(w)|^2}{|w|} dw < \infty \tag{3.5}$$

onde C_Ψ é a constante de admissibilidade e $\Psi(w)$ é a TF de $\psi(t)$. Para alcançar esta condição, basta que a média da função seja nula, isto é:

$$\int_{-\infty}^{+\infty} \psi(t) dt = 0 \tag{3.6}$$

o que implica que as wavelets devem ter um espetro do tipo passa-banda e uma forma semelhante a uma onda. Tais propriedades permitem uma localização eficaz em termos de

tempo e frequência, ao contrario da TF. A wavelet ψ é designada como a wavelet mãe, sendo apresentado um exemplo na figura 3.11.

Nas secções que se seguem serão feitas as apresentações dos dois tipos de TWs: a Transformada Continua (CWT) e a Discreta (DWT). Serão discutidos ambos os tipos de transformada, bem como os seus cenários de uso mais apropriados.

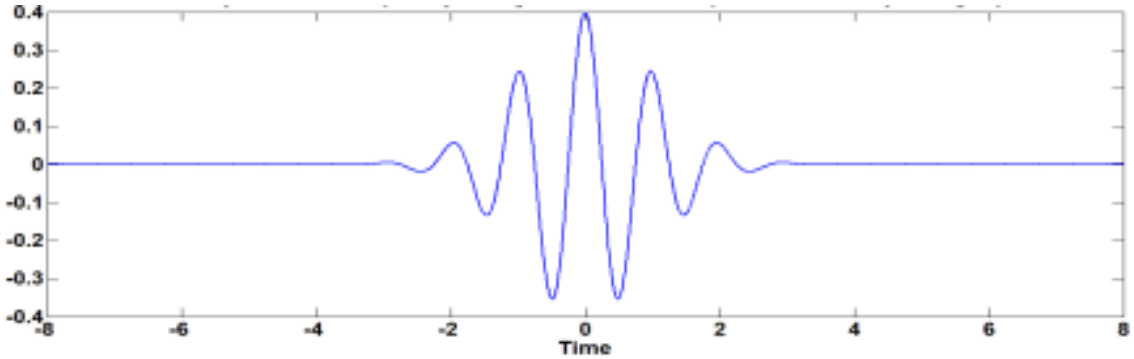


Figura 3.11: Wavelet típica.

Transformada de Wavelet Contínua

A Transformada de Wavelet Contínua (CWT) é definida a partir de um espaço de funções ortonormais que constituem a base das funções da TW.

O uso de uma decomposição wavelet baseada na CWT permite a análise de qualquer processo nos domínios do tempo e da frequência. Através de mudanças de escala e de translações, a wavelet mãe ψ dá origem a um conjunto de funções wavelet "filhas"$\psi_{(\tau,s)}$:

$$\psi_{(\tau,s)}(t) = \frac{1}{\sqrt{|s|}}\psi(\frac{t-\tau}{s}), \quad s,\tau \in \mathbb{R}, s \neq 0 \tag{3.7}$$

onde τ e s são os parâmetros de translação e de mudança de escala, respetivamente. O primeiro parâmetro é usado para o deslocamento da wavelet mãe no tempo, enquanto o segundo parâmetro é usado para o controlo da largura da janela de análise e, consequentemente, da frequência que está a ser analisada. Como se trata de uma transformada contínua, tanto o τ como o s devem ser incrementados continuamente e a transformada tem que ser integrada ao longo do tempo, o que torna o cálculo desta transformada uma tarefa computacionalmente pesada. Ao variar esses parâmetros, pode ser realizada uma análise multi-escalar de todo o processo, fornecendo uma descrição das diferentes componentes da frequência juntamente com os intervalos de tempo onde cada uma dessas componentes está localizada.

Dada uma série temporal $x(t) \in L^2\mathbb{R}$, a sua CWT $C_x^{\psi}(\tau, s)$ pode ser definida como [60]:

$$C_x^{\psi}(\tau,s) = \frac{1}{\sqrt{|s|}}\int_{-\infty}^{+\infty} x(t)\psi^*(\frac{t-\tau}{s})dt, \quad s,\tau \in \mathbb{R} \tag{3.8}$$

onde $*$ denota a conjugação complexa da função wavelet básica $\psi(t) \in L^2(\mathbb{R})$ e $\frac{1}{\sqrt{|s|}}$ é utilizada como um fator de preservação da energia.

Através da análise da série temporal original em toda a gama de escalas de decomposição, as CWTs são capazes de fornecer uma representação da série nos domínios do tempo e da frequência.

Um Escalograma pode ser definido como a representação da energia normalizada $\hat{E}_x(\tau, s)$ ao longo de todas as translações possíveis (conjunto $\mathbf{T}$) em todas as escalas analisadas (conjunto $\mathbf{S}$), e é calculada como:

$$\hat{E}_x(\tau, s) = 100 \frac{|\Psi_x^{\psi}(\tau, s)|^2}{\sum_{\tau' \in \mathbf{T}} \sum_{s' \in \mathbf{S}} |\Psi_x^{\psi}(\tau', s)|^2} \tag{3.9}$$

O volume delimitado pela superfície do escalograma é o valor quadrático da média do processo. A análise destes escalogramas permite a descoberta das diferentes componentes da frequência, para cada escala de análise. Por exemplo, a existência de um pico no escalograma a uma frequência baixa indica a existência de um componente de frequência baixa na análise das séries temporais, enquanto que um pico no escalograma a uma frequência alta corresponde a um componente de alta frequência. Além disso, partindo do principio que o processo $x(t)$ é estacionário ao longo do tempo, várias informações estatísticas podem ser obtidas, como o desvio padrão:

$$\sigma_{x,s} = \sqrt{\frac{1}{|\mathbf{T}|} \sum_{\tau \in \mathbf{T}} (\hat{E}_x(\tau, s) - \mu_{x,s})^2}, \forall s \in \mathbf{S} \tag{3.10}$$

onde $\mu_{x,s} = \frac{1}{|\mathbf{T}|} \sum_{\tau \in \mathbf{T}} (\hat{E}_x(\tau, s)$, e $|\mathbf{T}|$ representa a cardinalidade do conjunto $\mathbf{T}$.

Esta abordagem é utilizada nesta dissertação para efectuar a decomposição, análise e classificação do tráfego, uma vez que se pretende que tal tarefa seja realizada em tempo real. A CWT será intensivamente utilizada na metodologia de classificação de tráfego proposta no capítulo 4.

Transformada de Wavelet Discreta

A Transformada de Wavelet Discreta (DWT) também podem ser utilizada para representar as funções e os sinais no tempo e nas diversas componentes de frequência. Uma vantagem da DWT é o fato de ser computacionalmente menos complexa, uma vez que a sua complexidade é $O(N)$ enquanto que a complexidade da *Fast Fourier Transform* (FFT) é $O(Nlog(N))$. Realizando uma mudança de escala, que pode consistir numa expansão ou compressão, e um deslocamento temporal sobre a wavelet "mãe", obtemos $\psi_{j,k} = 2^{-j/2}\psi(2^{-j}t - k)$, ou seja, a frequência de oscilação central move-se para $2^{-j}f_0$ e a origem da referência temporal para $2^j k$. Note-se que j representa a escala temporal, k representa o k^{th} coeficiente correspondente à escala j, com j_0 representando a maior escala de tempo. A DWT também utiliza um filtro passa-baixo, $\phi(t)$, conhecido como função de escalonamento, que pode ser estendida e temporariamente deslocada de uma maneira similar à função $\psi(t)$. Portanto, um sinal $x(t)$ pode ser construído como a soma das funções de escala e de wavelet:

$$x(t) = \sum_k c_x(j_0, k)\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty} \sum_k d_x(j, k)\psi_{j,k}(t) \tag{3.11}$$

onde $\phi_{j_0,k}(t)$ e $\psi_{j,k}(t)$ são, respectivamente, a função de escala genérica e a função wavelet genérica. $c_x(j_0,k)$ são os coeficientes de escala e $d_x(j,k)$ são os coeficientes da wavelet. O logaritmo dos coeficientes da wavelet, para o momento de ordem q, pode ser definido como:

$$y_{q,j} = \log_2(\frac{1}{K}\sum_{k=1}^{K}|d_x(j,k)|^q), q \in \mathbb{R} \tag{3.12}$$

que serão doravante designados genericamente como estimadores multi-escalar, em que K é o número de coeficientes a analisar na escala de temporal j. O comportamento da escala de qualquer processo estocástico pode ser estudado através de uma análise do diagrama *Log-scale*, que é um gráfico log-log dos estimadores $y_{q,j}$ dos detalhes da wavelet em cada escala, em função da escala, completado com o intervalo de confiança das estimativas de cada escala [61].

3.2.3 Análise Multi-escalar do tráfego

Nesta dissertação vamos utilizar wavelets para decompor o tráfego de rede em várias escalas, ou seja, diferentes níveis de agregação, com o objetivo de avaliar e correlacionar as diferentes características das componentes do espetro da frequência que correspondem aos diversos mecanismos subjacentes da rede. Como explicado na secção 3.1, o tráfego da Internet é gerado por eventos de baixa frequência, como as solicitações dos utilizadores, e controlado por mecanismos presentes nas componentes de frequência intermédias do espetro, como as diferentes sessões de Internet e os diferentes mecanismos de controlo de tráfego existentes. Todos estes eventos e mecanismos criam eventos de alta frequência que correspondem à chegada de pacotes. Por exemplo, quando um utilizador faz uma solicitação utilizando uma aplicação da Internet, como por exemplo clicar num *link* de um *website* ou solicitar um vídeo *on-line*, diversos processos são criados pelo sistema operativo. Cada um desses processos cria um conjunto de sessões de Internet, cada uma gerando um fluxo de tráfego. Ao nível da camada de rede, cada uma destas ligações irá transmitir e receber os dados solicitados em vários pacotes. A análise de cada mecanismo pode ser avaliada através de escalas de agregação apropriadas, ou escalas de frequência. Isto é ilustrado na figura 3.12, que representa a forma como os mecanismos presentes nas diferentes escalas estão relacionados, como formam o tráfego de Internet e como podem ser analisados. Ao analisar o tráfego gerado por uma aplicação, mostrado no lado esquerdo da figura 3.12, e fazendo um *zoom* das dinâmicas observadas, somos capazes de inferir todos os mecanismos presentes nas diferentes escalas de análise e avaliar a sua influência na dinâmica global do tráfego. Componentes como os intervalos de tempo entre os pedidos do utilizador (representados por Δ_1) ou os instantes iniciais (representado por Δ_{2x}) podem ser avaliados através da realização de uma mudança na escala de análise, o que corresponde à realização de um *"zoom in"* no tráfego analisado.

O conceito principal da nossa abordagem consiste na avaliação da presença de cada um dos mecanismos mencionados, o que pode ser feito usando a escala de agregação ou escala de frequência apropriada. Desta forma, será possível obter assinaturas espetrais características que descrevam as várias componentes de frequência para cada aplicação estudada, permitindo a discriminação precisa das diversas aplicações.

O conceito subjacente a esta abordagem consiste em analisar as diversas interações criadas por uma aplicação de Internet. Estas podem consistir em várias sessões simultâneas com diferentes hosts e servidores remotos, tal como mostra a figura 3.4. Para o tráfego não cifrado, o tráfego das diferentes aplicações pode ser monitorizado separadamente, enquanto

que para o tráfego cifrado isto pode não ser possível, como é ilustrado na figura 3.20. Por isso, foi proposta a definição de *data-stream*, apresentada na seção 3.1.1, no sentido de representar todo o tráfego que é enviado e recebido por uma classe de aplicação de Internet e é identificado por um identificador numérico. Esse identificador pode incluir (*i*) a porta local/remota, no caso de tráfego não cifrado, ou (*ii*) qualquer identificador específico da tecnologia de túnel cifrado, no caso de tráfego cifrado.

A análise dos *data-streams* ao longo do comprimento $\Delta(t)$ pré-definido, como é ilustrado na 3.20, permite uma monitorização contínuo do tráfego de cada aplicação, aumentando a precisão de classificação.

A análise multi-escalar baseada na CWT pode também ser utilizada nas abordagens de profiling de utilizadores.

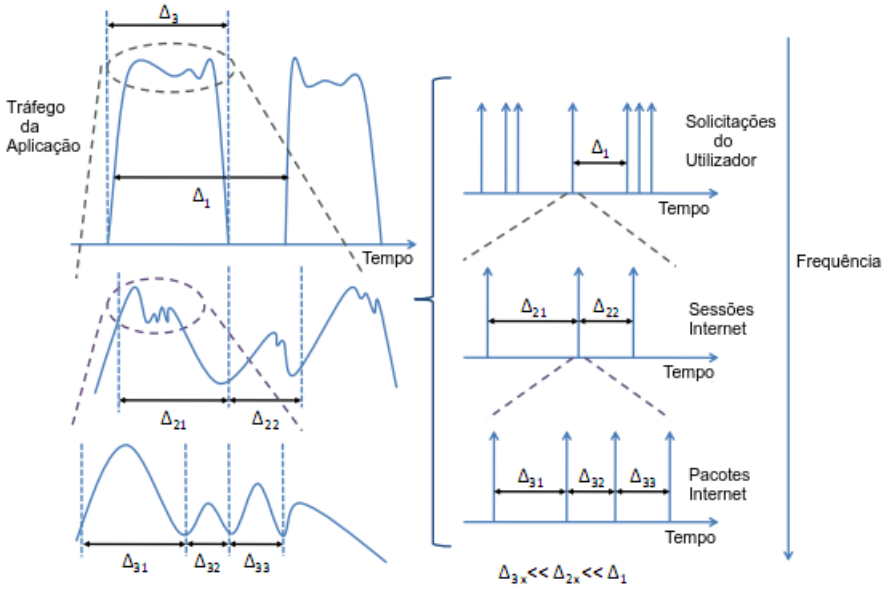


Figura 3.12: Dinâmica do Tráfego Multi-Escalar [editado de [3]].

Escalogramas Wavelet para a análise Multi-Escalar

Como foi dito anteriormente,a incapacidade da TF convencional para preservar a dependência do tempo e descrever as características espetrais evolutivas dos processos não estacionários requer ferramentas que permitem uma localização simultânea no tempo e frequência. As TWs podem fornecer essa informação, permitindo que as componentes locais, transitória ou intermitentes sejam realçadas. Essas componentes são muitas vezes ocultadas pela média utilizada nos diversos métodos espetrais.

As Wavelets permitem a análise de cada uma das componentes de um sinal numa escala apropriada. Fazendo a analogia com a terminologia utilizada na TF, o espetro de energia da wavelet (local), também chamado de Escalograma ou Espetrograma, é definido de acordo com a fórmula 3.9.

Os escalogramas possibilitam a descoberta de diferentes componentes de frequência, para

cada escala de análise. Como afirmado anteriormente, a existência de um pico no escalograma para uma frequência baixa indica a existência de um componente de baixa frequência na série temporal analisada, enquanto que um pico no escalograma numa frequência alta corresponde a uma componente de alta frequência. Os escalogramas revelam muitas informações sobre a natureza dos processos não estacionários, que podem ser úteis em várias áreas científicas: diagnósticos de eventos especiais em comportamento estrutural durante um terramoto, análise do movimento do solo, análise da resposta de edifícios perante tempestades, entre outros [62].

Análise do escalograma de tráfego

Nesta dissertação será efetuada uma decomposição wavelet, recorrendo à CWT, das métricas do tráfego capturado ao nível da camada 2. Os escalogramas obtidos serão normalizados ao longo do comprimento do processo, tal como descrito na equação 3.9. As figuras 3.13 até 3.18 mostram as métricas de tráfego, a taxa de *download* em bytes por segundo (amostrados em intervalos de 0.1 segundos) e os escalogramas correspondentes ás diferentes aplicações *web* que foram apresentadas no capítulo 3.1. A análise destas figuras revela caraterísticas diferentes que são causadas pelos padrões de tráfego distintos apresentados pelas aplicações, e que têm origem nas caraterísticas das interações humanas e das redes/serviços.

As aplicações de redes sociais *on-line* (Figura 3.13) geram tráfego que apresenta picos mais frequentes e de menor amplitude, que são gerados pelas atualizações de estado criadas pelas ligações de outros utilizadores, que normalmente consistem apenas em mensagens. Há portanto menos componentes de baixa frequência, enquanto que as componentes de alta frequência estão menos presentes no processo devido à pequena quantidade de tráfego trocado.

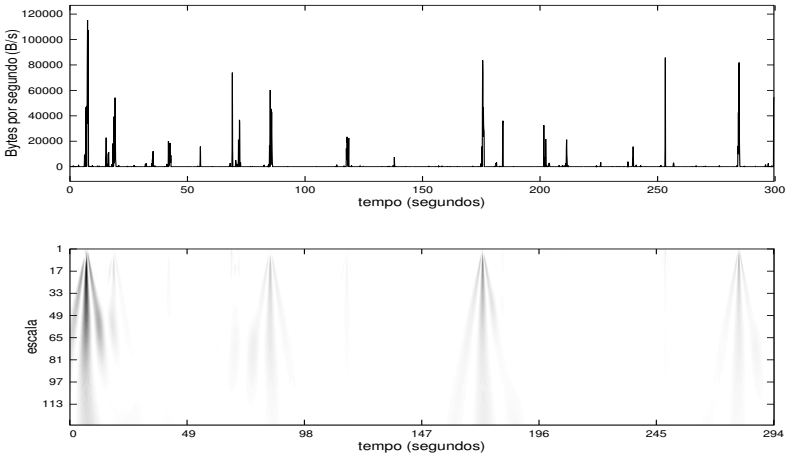
O tráfego de notícias *on-line* (Figura 3.14) apresenta vários picos aperiódicos de curta duração e amplitude considerável. Estes picos são causados pelos cliques dos utilizadores em *hiperlinks*, enquanto navegam pelas notícias disponíveis, causando o *download* de uma nova página que apresenta as notícias solicitadas e criando componentes consideráveis de baixa frequência. Além disso, os escalogramas gerados por estas aplicações apresentam alguns componentes consideráveis de frequência média devido ao número considerável de sessões TCP criadas, existindo algumas componentes consideráveis de alta frequência relacionadas com a chegada dos pacotes.

As aplicações de *email on-line* (figura 3.15) geram tráfego com poucos picos e pouco frequentes, que correspondem à sincronização inicial e automática entre o servidor e o cliente. Estes picos são de curta duração e menos frequentes do que os das aplicações *on-line* apresentadas anteriormente. No entanto, há pequenas componentes de alta frequência causadas pelo tráfego de sincronização que raramente procura por novos *e-mails*, enquanto que os componentes de baixa frequência não estão muitos espalhadas pelo escalograma de tráfego devido a natureza periódica dos eventos de rede/serviços.

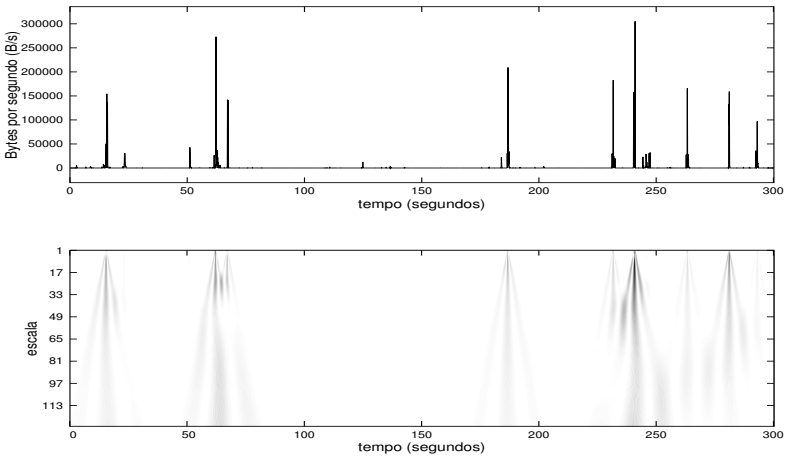
Aplicações de partilha de fotos *on-line* (Figura 3.16) geralmente geram vários picos de tráfego com pseudo-periodicidade, devido aos cliques que são executados pelo utilizador enquanto pede para ver outra fotografia. Estes picos são geralmente de baixa amplitude, já que consistem no *download* de uma fotografia utilizando uma única sessão TCP. Consequentemente, podem-se observar várias componentes de alta frequência e de baixa amplitude espalhadas

por todo o escalograma, existindo também algumas componentes de baixa frequência.

Por último, serviços de vídeo *on-line* (Figuras 3.17 e 3.18) geram tráfego com intervalos pequenos entre chegadas de pacotes, causados pelo *download* do vídeo solicitado usando a largura de banda da rede disponível. Consequentemente, existem componentes consideráveis de alta frequência causadas pela chegada dos pacotes, não existindo componentes significativas de baixa frequência já que não há muitos cliques por parte dos utilizadores.

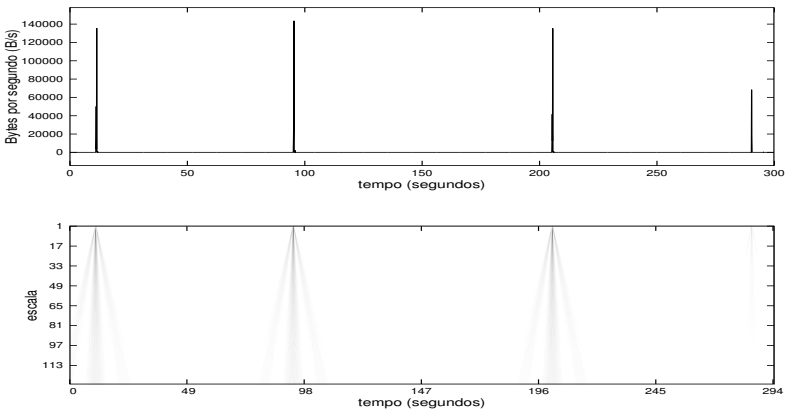


(a) Tráfego Facebook - direção download.

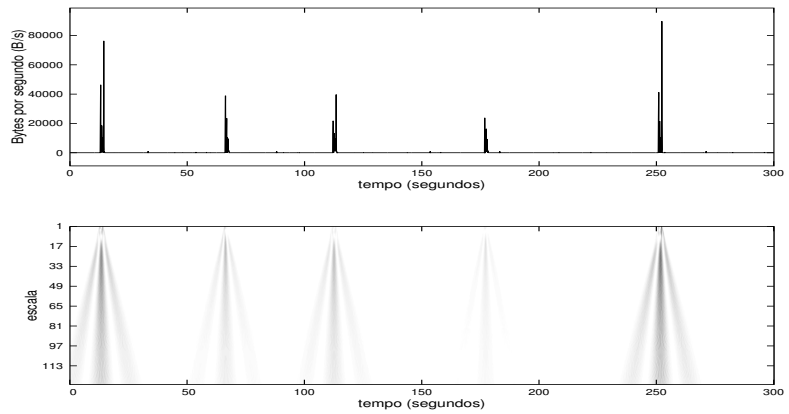


(b) Tráfego Google^{+} - direção download.

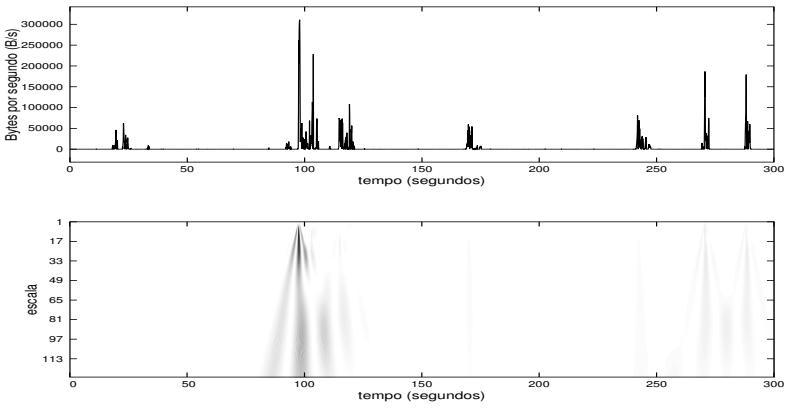
Figura 3.13: Padrões de Tráfego das Rede Sociais *On-line* Facebook 3.13a e Google^{+} 3.13b e Escalogramas correspondentes.



(a) Tráfego Abola - direção download.

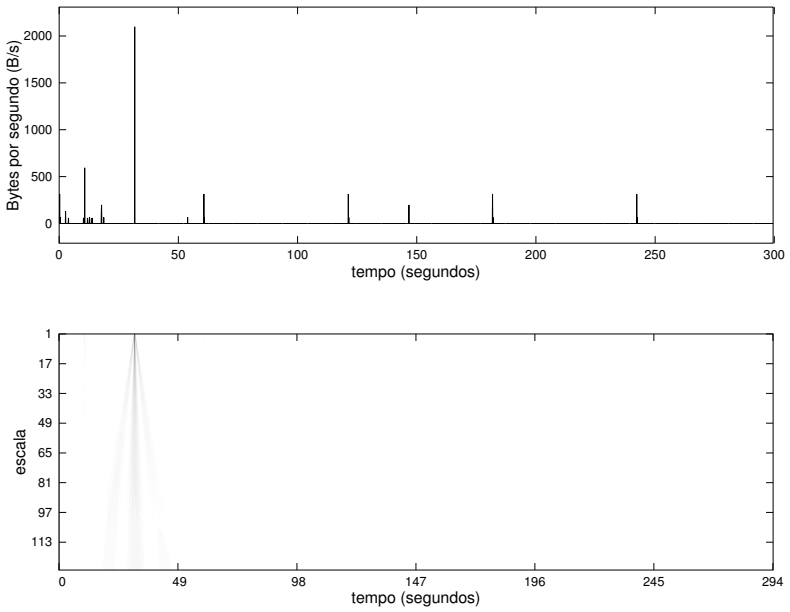


(b) Tráfego Record - direção download.

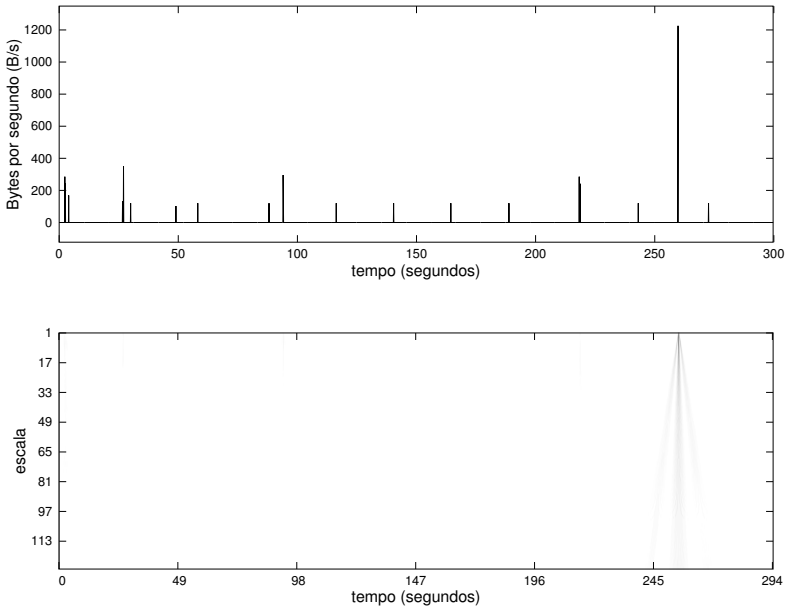


(c) Tráfego CNN - direção download.

Figura 3.14: Padrões de Tráfego dos Serviços de Notícias *On-line* Abola 3.14a, Record 3.14b e CNN 3.14c e Escalogramas correspondentes.

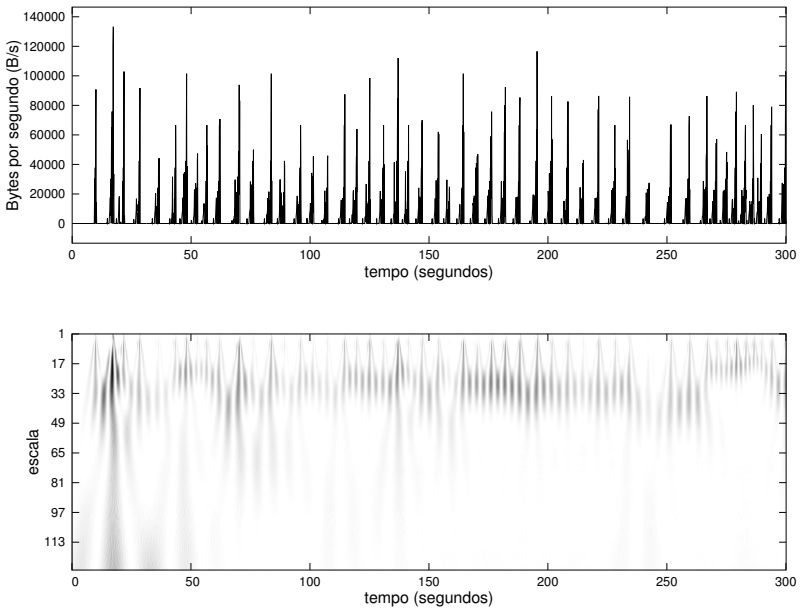


(a) Tráfego Hotmail - direção download.

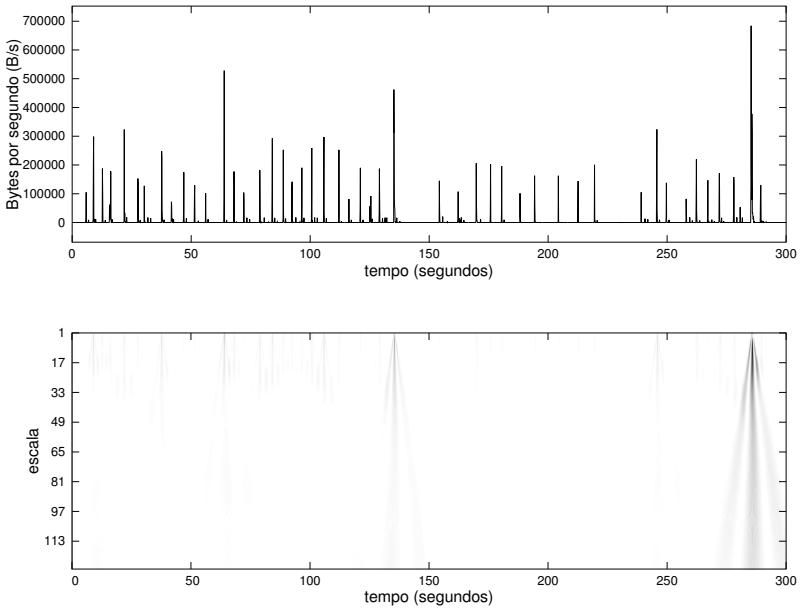


(b) Tráfego Gmail - direção download.

Figura 3.15: Padrões de Tráfego dos Serviços *E-mails On-line* Hotmail 3.15a e Gmail 3.15b e Escalogramas correspondentes.

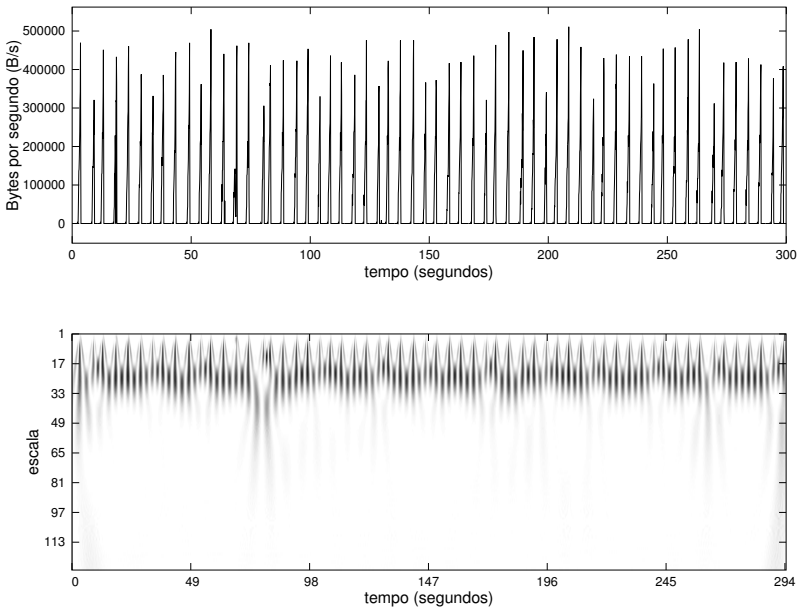


(a) Tráfego Flickr - direção download.

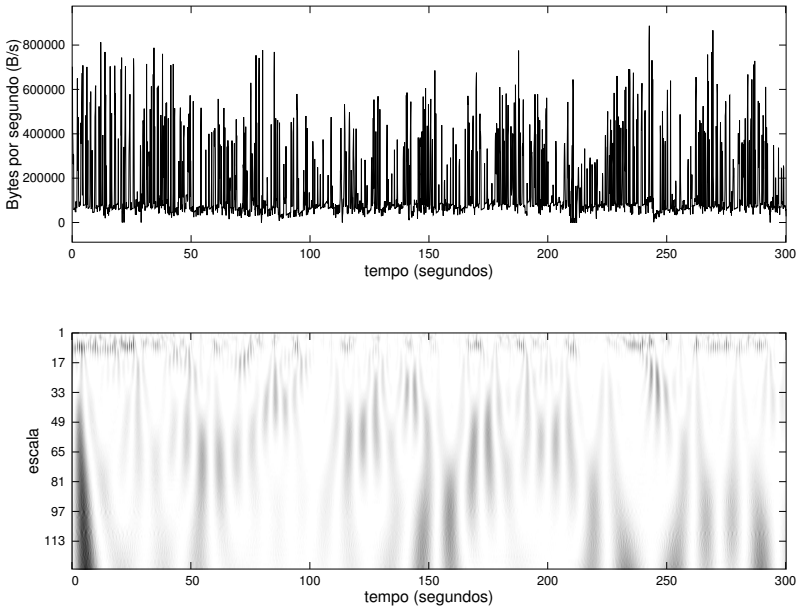


(b) Tráfego Picasa - direção download.

Figura 3.16: Padrões de Tráfego de Serviços de Partilha de Fotos *On-line* Flickr 3.16a e Picasa 3.16b e Escalogramas correspondentes.

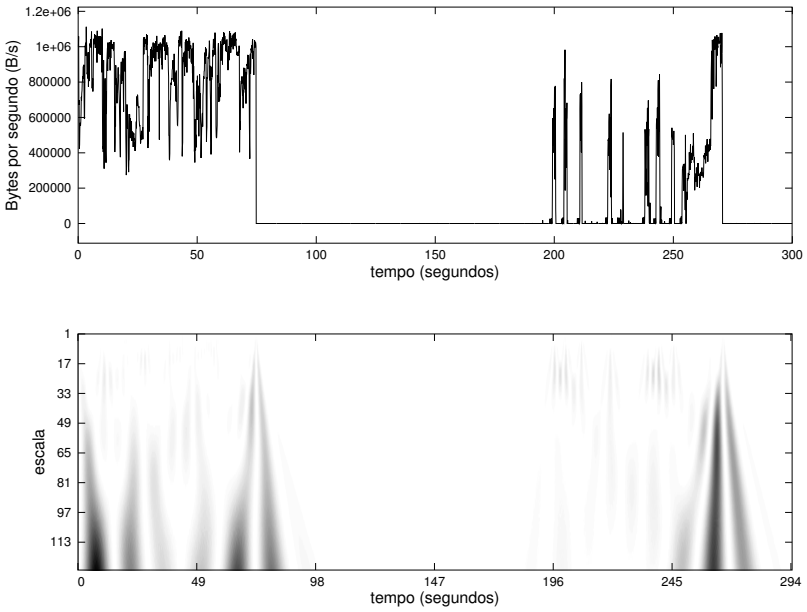


(a) Tráfego Youtube - direção download.

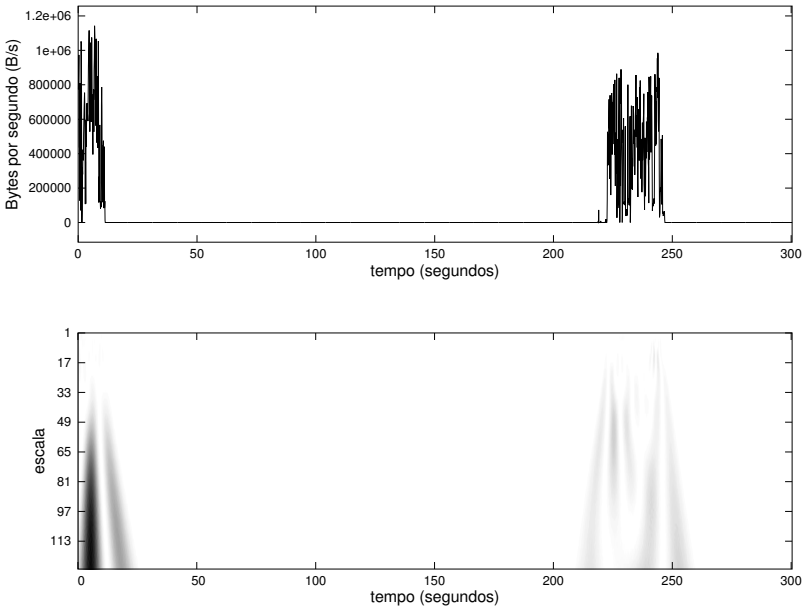


(b) Tráfego vimeo - direção download.

Figura 3.17: Padrões de Tráfego de Serviços de Partilha de Vídeos HD *On-line* (VLD) Youtube 3.17a e vimeo 3.17b e Escalogramas correspondentes.



(a) Tráfego Youtube - direção download.



(b) Tráfego vimeo - direção download.

Figura 3.18: Padrões de Tráfego de Serviços de Partilha de Vídeos HD *On-line* (VCD) Youtube 3.18a e vimeo 3.18b e Escalogramas correspondentes.

Desvio padrão

Desvio padrão é uma medida de dispersão dos valores de uma distribuição em relação à média.

O desvio padrão s de um vetor de dados x pode ser definido como:

$$s = \sqrt{\frac{\sum (x_i - Media)^2}{n-1}} \tag{3.13}$$

onde $Media = \frac{1}{n}\sum_{i=1}^{n} x_i$ e n é o numero dos elementos da amostra.

O cálculo do desvio padrão é um aspeto de crucial importância no nosso trabalho, porque calculando o desvio padrão dos coeficientes de wavelet dos diferentes *data-streams* (escolhidos aleatoriamente do conjunto de dados) pertencentes a cada aplicação *web* será possível identificar o comportamento de cada aplicação, como mostra a figura 3.19. De facto, ao analisar o perfil de variação da energia do processo em toda a gama de frequências é possível obter uma associação precisa entre um dado fluxo de tráfego e a aplicação que lhe deu origem através da análise das regiões diferenciadoras e usando os algoritmos descritos na seção 4 para avaliar a precisão da classificação das diferentes aplicações *web*.

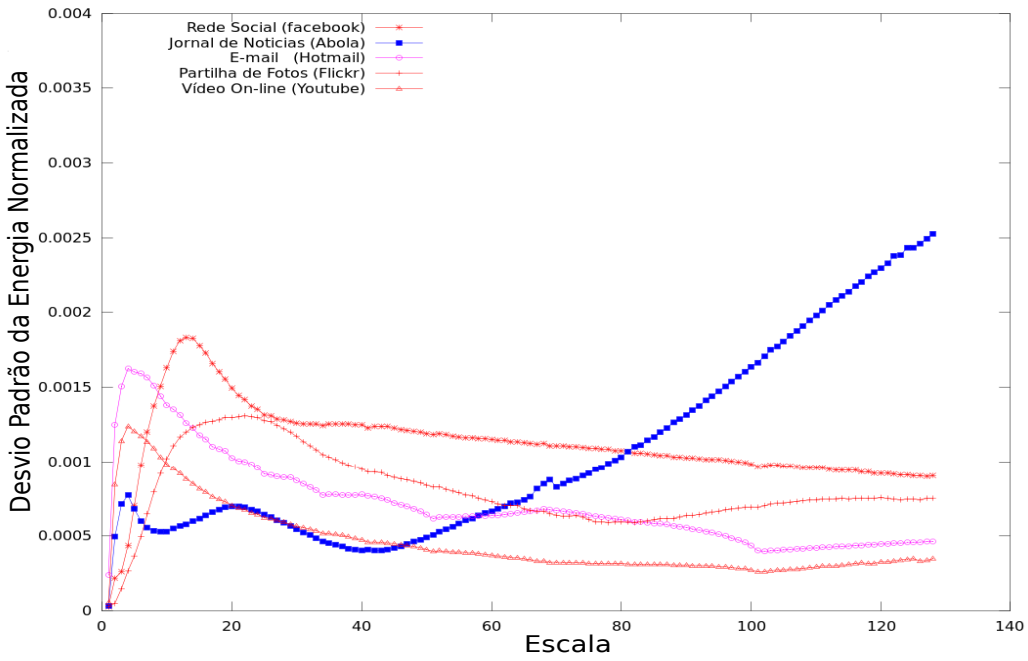


Figura 3.19: Desvio padrão da escala de análise dos fluxos de cada aplicação.

3.2.4 Definições Preliminares

Nesta sub-seção serão apresentadas algumas definições importantes para a abordagem de classificação que é utilizada. Nesta dissertação os *Traces* de tráfego são divididos em dois tipos. Os *traces* de treino consistem no conjunto de dados estatísticos retirados dos *data-streams* pertencentes a tráfego que foi identificado através da técnica DPI e que foi gerado em ambiente laboral controlado. Estes *traces* são utilizados para identificar as classes de classificação e para inferir os parâmetros das diferentes distribuições probabilísticas. O segundo tipo é constituído por *traces* de teste, que são utilizados para avaliar a precisão da metodologia proposta.

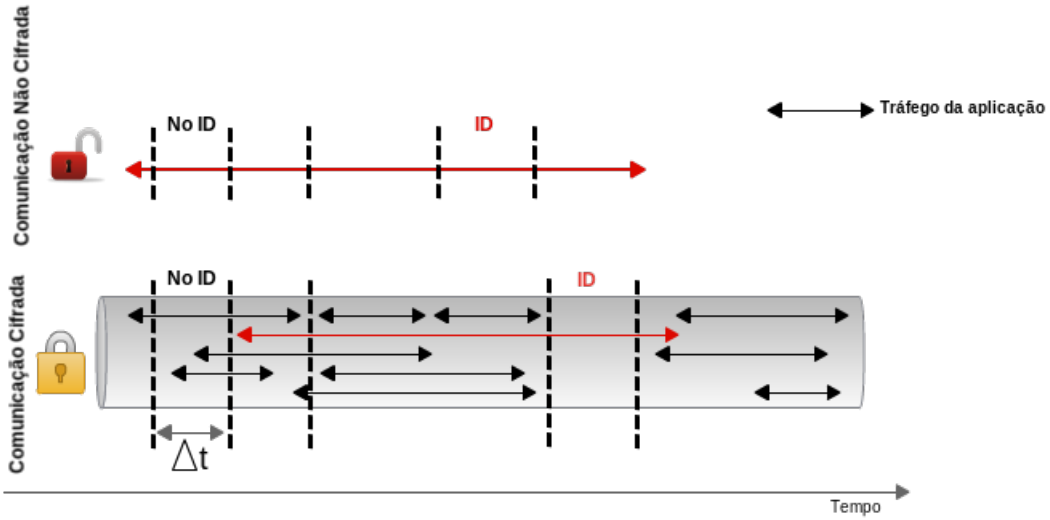


Figura 3.20: Conceito de Classificação do Tráfego [editado de [3]].

Consideremos que A representa o número de aplicações estudadas, P representa o número de *traces* de treino de cada aplicação Internet que será classificada, N representa o número de pontos dos *traces* de teste, Z é o número de regiões ao longo do gráfico do desvio padrão (Figura 3.19) e T representa o número de escalas ao longo do tempo de cada CWT analisada. Seja ainda

$$D_z = (a, p, t), a = 1, ..., A, p = 1, ..., P, t = 1, ..., T \tag{3.14}$$

o $z^{ésimo}$ elemento do conjunto que indexa todos os processos estocásticos que são extraídos, em cada escala, a partir de todos os *data-streams* de cada *trace* de treino, de cada aplicação, ao longo de todas as regiões z. Um conjunto análogo pode ser definido para os *traces* de teste.

Na figura 3.19 os *traces* de teste estão identificados com a cor azul, enquanto que os *traces* de treino são identificados com outras cores.

3.3 Parâmetros de Validação

No sentido de avaliar a eficiência do método, são utilizados alguns parâmetros de validação. Comecemos por apresentar algumas definições importantes.

Para uma categoria de aplicações A, designemos por:

- **Verdadeiro Positivo (VP)** - Os fluxos da aplicação A que são classificados corretamente como pertencentes a A.

- **Falso Positivo (FP)** - Os fluxos que não pertencem a A e que são classificados corretamente como não pertencentes a A.

A precisão é um dos parâmetros de validação que serão utilizados:

1. **Precisão** - Percentagem de amostras classificadas como A e que são realmente pertencentes à classe A, a qual pode ser calculada por $VP/(VP + FP)$. A precisão pode ser atribuída a uma classe especifica ou a várias classes, em média.

A *matriz de classificação* será também utilizada para descrever os detalhes dos resultados da classificação: **Matriz de Classificação** - Matriz N x N, cujo elemento (i,j) é a percentagem dos fluxos que pertencem à classe i e que são classificados como pertencentes à classe j. Esta matriz mostra os detalhes dos resultados da classificação, que podem ser utilizados para explorar as características inerentes de cada aplicação.

3.4 Cenários de recolha e processamento de tráfego

Nesta sub-seção é apresentado o cenário de medição, os meios utilizados nos testes e as condições em que foram efetuadas as capturas. Mostra-se ainda como foram obtidos os dados para os vários casos estudados, apresentando-se ainda os programas utilizados.

3.4.1 Cenário de medição

Todos os testes inseridos nesta dissertação foram realizadas no laboratório de redes do Instituto de Telecomunicações da Universidade de Aveiro. A máquina utilizada para a realização dos testes é um computador pessoal com processador Intel Core 2 Duo E8500 com 3 GB de memória RAM DDR 3.

Foi utilizado o sistema operativo Ubunto Desktop Edition, versão 10.04, numa primeira fase, enquanto que na fase final se utilizou a versão 12.04 para a realização dos testes, fazendo uso do web browser Mozila Firefox para aceder ás aplicações *Web*.

3.4.2 Monitorização e Captura do Tráfego

As principais ferramentas hoje existentes para efetuar a monitorização e coleta de dados numa rede TCP/IP são:

1. **Tcpdump e Libpcap**

   O programa Tcpdump, disponível em [63], é um analisador de tráfego desenvolvido em código aberto utilizando a linguagem de programação "C", estando disponível para os sistemas operativos UNIX/Linux. Tcpdump realiza a captura e filtragem de pacotes que passam pela rede através de uma interface de rede colocada em modo promíscuo, ou seja, a interface "escuta"todo o tráfego da rede à qual está ligada. O uso de várias

opções de filtragem permite selecionar os tipos de protocolos de rede, endereço origem, endereço destino, porta origem, porta destino, etc. Este programa foi desenvolvido a partir de uma biblioteca de funções para captura de pacotes chamada *Library Packet Capture* (Libpcap), que também foi desenvolvida utilizando a linguagem "C".

O Tcpdump não tem uma interface gráfica, é executado em linha de comando num ambiente UNIX/Linux e fornece como saída informações em texto puro. O programa permite gravar em ficheiro as informações das capturas dos pacotes de rede, possibilitando uma análise do ficheiro através da utilização de outras ferramentas. Há também uma versão do tcpdump , o Windump, disponível em [64], que utiliza a Winpcap, que é uma libpcap para o Windows. Libpcap é a biblioteca utilizado pelo Tcpdump no processo de captura dos pacotes a nível do utilizador.

2. **Wireshark**

   O Wireshark [55] é um popular analisador de pacotes de rede. Inicialmente designado por Ethereal, o projeto foi renomeado em maio de 2006 como Wireshark devido a problemas de marca registada. É um software multi-plataforma e ao contrário do Tcpdump proporciona ao utilizador uma interface gráfica denominada de *GIMP Toolkit* (GTK) (*GNU Image Manipulation Program* (GIMP)), e utiliza a biblioteca pcap (Libpcap) na captura dos pacotes que transitam pela rede. Pode ser executado em vários sistemas operativos baseados em Unix, incluindo Linux, Mac, OS X, BSD e Solaris, e também em Windows. Existe uma versão baseada em terminal, ou seja, não dispondo de interface gráfica, chamada Tshark [65].

   O Wireshark é muito parecido com o tcpdump, mas possui uma interface gráfica que facilita a iteração do utilizador com as opções de captura, e apresenta muito mais opções de classificação e filtragem de informação. O Wireshark fornece ao utilizador várias opções de configuração do modo de captura do tráfego, e permite que o utilizador visualize todo o tráfego da rede em tempo real.

   Tal como o Tcpdump, o Wireshark faz uso da biblioteca windpcap e do formato de arquivo em libpcap.

Foi feita a captura do tráfego TCP, HTTP e HTTPS (para o *site* Facebook) em modo root, utilizando a ferramenta Wireshark disponível para o Linux e com o modo promíscuo desligado. As capturas obtidas foram posteriormente transformadas em tabelas com dados referentes ao número de pacotes e número total de bytes por intervalo de tempo utilizando as opções estatísticas do Tshark. Posteriormente, foram criados gráficos referentes aos dados capturados utilizando o Octave, de forma a facilitar a leitura dos valores e tentar reconhecer padrões no tráfego capturado para cada *site*.

3.4.3 Processamento e amostragem de tráfego

Uma vez que o objetivo do estudo consiste em observar a forma como o tráfego é transmitido e se existem padrões na transmissão que identifiquem o tipo de fluxo, o processamento necessário é mínimo, consistindo apenas em tratar os dados capturados através da componente estatística do programa Tshark, disponível em [65]. Os dados foram processados utilizando o comando "Tshark -r ficheiro.pcap -q -z io, stat, 0.1, ip.dst==193.93.69.26,

ip.src==193.93.69.26", onde foi feita uma divisão entre o tráfego recebido (*download*) e o tráfego enviado (*upload*) através do uso do filtro do Tshark, utilizando a opção "ip.src=="para que se obtenha o tráfego enviado e "ip.dst=="para que se obtenha o tráfego recebido, em que "193.93.69.26"é o IP da máquina de testes. Partindo da divisão do tráfego, são obtidos vários ficheiros com o tráfego recebido e enviado a partir das diversas aplicações *Web*.

3.4.4 Visualização e análise de tráfego

Após o processamento dos dados, é necessário efetuar a sua análise. Para isso, são criados gráficos onde os dados são representados para uma mais fácil observação e posterior análise. Os gráficos foram criados recorrendo ao programa Octave [66], em detrimento do MatLab [67], já que possibilita uma criação mais personalizada dos gráficos. Desta forma, foi possível dar aos dados uma visibilidade diferente, aproveitando também a maior resolução oferecida pelas imagens dos gráficos.

Apesar de terem sido usados diferentes intervalos de tempo, consoante o tipo de tráfego e a duração da captura, optou-se por representar sempre o tráfego em bps em todos os gráficos de forma a facilitar a perceção dos dados e as comparações entre os diversos tipos de tráfego.

3.5 Sumário

Neste capítulo foram abordados vários conceitos e definições que são intensamente utilizados nos capítulos seguintes. Inicialmente foi fornecida uma explicação sobre a dinâmica do tráfego gerado pelas diferentes aplicações da Internet. Em seguida, foi apresentada a definição de *data-stream*, explicando como o tráfego pode ser agrupado em *data-streams* de acordo com as restrições de classificação. Foram também apresentadas as diferentes aplicações de Internet que foram estudadas, e o respetivo tráfego. Foram dadas algumas noções sobre análise escalar de tráfego, abordando algumas das metodologias mais relevantes existentes para a análise das diferentes comnponentes de um sinal. Foram discutidas as vantagens e os problemas associados a cada abordagem, seguindo-se a apresentação das diferentes metodologias de análise multi-escalar de tráfego propostas nesta dissertação. Foram ainda apresentadas algumas definições preliminares, assim como os parâmetros de validação que serão utilizados. Por fim, foram apresentados os cenários de recolha e processamento de tráfego.

Capítulo 4

Metodologia Proposta para caraterização do tráfego

Nesse capítulo será proposta uma metodologia para a classificação/identificação de aplicações *web*, e que se baseia no cálculo do peso de cada *trace* de uma aplicação específica ao longo de diversas regiões do seu escalograma. Serão propostos dois algoritmos diferentes para o cálculo do peso. Têm sido propostas diversas abordagens de classificação, embora ainda não exista um consenso na literatura sobre qual é o melhor procedimento. O intuito da metodologia proposta é utilizar um método que não apresente os problemas dos métodos mencionados no capítulo 2 e que seja simultaneamente funcional. Através da analise dos componentes das escalas dos *data-streams* de cada aplicação da Internet, será possível construír classes representativas das diferentes aplicações e identificá-las.

A metodologia proposta classifica as aplicações tendo em conta duas variáveis: o tamanho e a duração dos *traces* de tráfego. O procedimento utilizado é o mesmo para cada uma das variáveis e possibilita a classificação dos *data-streams* em N classes, onde N deve ser escolhido de acordo com as necessidades e características de cada rede. Para a aplicação dos algoritmos de classificação é necessário que os *data-streams* sejam recolhidos e armazenados num banco de dados contendo as informações relativas a cada *data-stream*: tamanho em bytes, duração em segundos, IP de destino e número da porta de destino.

4.1 Arquitetura do Sistema e Metodologia de Classificação

4.1.1 Visão Geral do Sistema

Conforme mencionado no capítulo 1, o sistema proposto consiste na análise do tráfego capturado referente a um determinado utilizador ligado a uma rede *wired*. A figura 4.1 mostra um diagrama que descreve a arquitetura geral e os diferentes componentes da abordagem de classificação proposta.

Para começar, o tráfego enviado a cada cliente é capturado, extraíndo-se diversas métricas da camada 2 (número de bytes e pacotes capturados). Será assim possível construir um perfil do tráfego de cada nó. As métricas da camada 2 extraídas são processadas através de uma análise multi-escalar feita através de uma decomposição baseada em wavelets (TW), utilizando a CWT, que constrói um perfil de tráfego multi-escalar baseando no escalograma inferido para alguns processos estatísticos escolhidos dos *traces* de tráfego capturados,

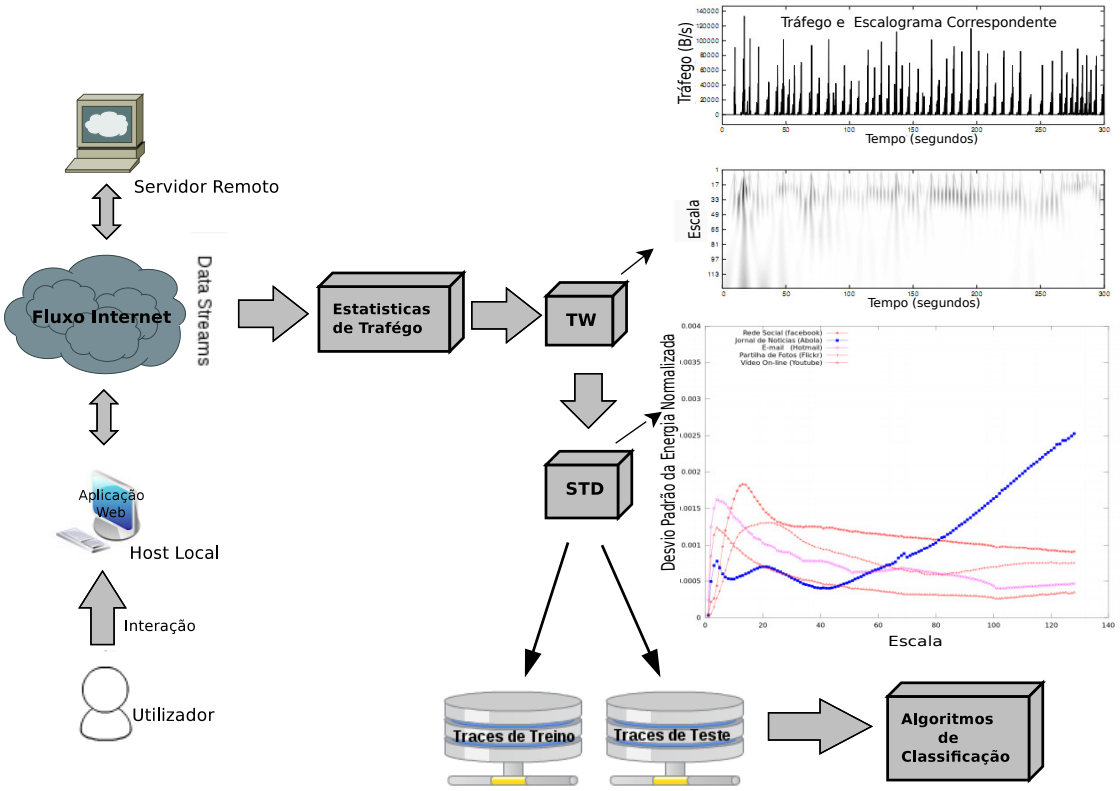


Figura 4.1: Arquitetura do Sistema.

conforme apresentado na seção 3.2.

Tais escalogramas descrevem as várias componentes de frequência presentes nas métricas do tráfego capturado, permitindo a associação das componentes ao evento do utilizador/rede correspondente. Inspecionando as diferentes componentes de frequência é possível inferir a preponderância no tráfego capturado dos vários eventos dos utilizadores/rede e identificar os perfis característicos das aplicações *Web*, permitindo assim classificar o tráfego capturado. Ao analisar parâmetros estatísticos, com o desvio padrão (*Standard Deviation* (STD)), dos diferentes escalogramas para cada escala de análise pode-se inferir a variabilidade da energia do processo e identificar as componentes de frequência mais proeminentes. Essa análise ajudará a encontrar componentes de frequência de diferenciação, permitindo classificar com precisão o tráfego. Os *traces* das várias aplicações são armazenados numa base de dados. Ao definir regiões caraterísticas das estatísticas dos escalogramas para os diferentes *traces* de treino das aplicações *Web*, é calculando o peso de cada aplicação ao longo das diversas regiões utilizando *traces* de teste, tornando possível identificar e caracterizar com precisão a aplicação *Web*.

São utilizados dois algoritmos para calcular o peso de um *trace* de uma determinada aplicação ao longo das diversas regiões.

4.1.2 Algoritmo 1

O cálculo do peso de um *trace* de uma determinada aplicação *web* (W_a), que define a precisão da classificação, é realizado através do somatório ao longo das diversas regiões ($\sum_{z=1}^{Z}$) da probabilidade do *trace* de treino da aplicação a ($P_z^{(a)}$) multiplicada pelo número de pontos

do *trace* de teste (N_z) ao longo das regiões, sendo por fim feito a divisão entre o resultado do somatório e o número total de pontos do *trace* de teste nas diversas zonas, como mostra a equação 4.1.

$$W_a = \frac{1}{N_{tz}} \sum_{z=1}^{Z} P_z^{(a)} N_z \tag{4.1}$$

4.1.3 Algoritmo 2

O cálculo do peso de um *trace* de uma determinada aplicação *web* (W_a), que define a precisão da classificação, é realizado através do produtório ao longo das diversas regiões ($\prod_{z=1}^{Z}$) da probabilidade do *trace* de treino de uma determinada aplicação ($P_z^{(a)}$) elevada ao número de pontos do *trace* de teste ao longo das diversas regiões (N_z), tal como mostra a equação 4.2.

$$W_a = \prod_{z=1}^{Z} P_z^{(a)N_z} \tag{4.2}$$

em que

a=1,2,3,4,5 se refere aos *traces* de treino das aplicações estudadas, concretamente facebook, abola, hotmail, flickr e youtube

Z é número máximo de zonas

W_a é peso de cada aplicação(1,2,3,4,5) em todas as zonas

N é número de pontos referente ao *trace* de teste

N_{tz} é número total de pontos dos *traces* de teste em todas as zonas

N_z é número de pontos dos *traces* de teste nas diversas zonas

$P_z^{(a)}$ Probabilidade de cada *trace* de treino referente a cada aplicação nas diversas zonas, com

$$P_z^{(a)} = \frac{N_z^{(a)}}{\sum_{a=1}^{A} N_z^{(a)}}$$

A é número máximo das aplicações = 5 (1,2,3,4,5)

$N_z^{(a)}$ Número de pontos dos *traces* de teste de cada aplicação nas diversas zonas

O cálculo do numero de pontos de um *trace* de uma determinada aplicação ao longo das regiões (ver figura 4.2) é efetuado mediante um varrimento dos pontos ao longo dos eixos x (escala de tempo ($MAXX$)) e y (desvio padrão da energia normalizada ($MAXY$)). Nesse sentido, são definidas variáveis de deslocamento ao longo do eixo x (d_x) e y (d_y), sendo α o fator que define a forma como é efetuado o deslocamento, tendo em conta as larguras w_x e w_y pré-definidas.

O deslocamento efetuado ao longo das regiões é calculado tendo em conta a largura (w_x e w_y) e o valor de α: $d_x = \alpha * w_x$ e $d_y = \alpha * w_y$. Este deslocamento permite determinar o nivel de precisão da classificação.

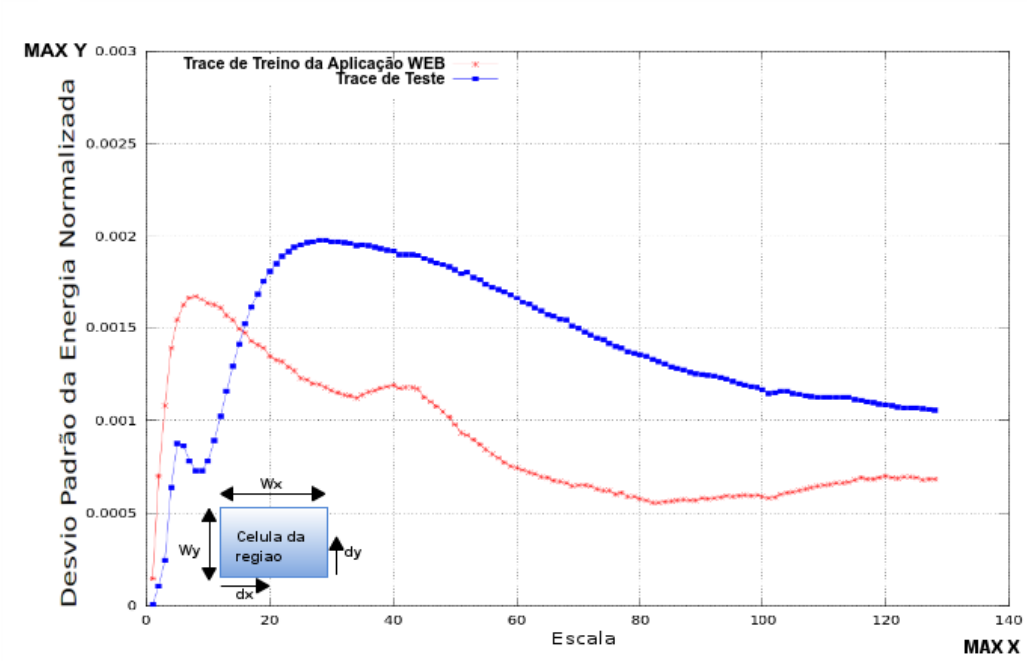


Figura 4.2: Cálculo do número de pontos de um *trace* de uma determinada aplicação.

4.2 Validação do método

Um *trace* F pertence a uma determinada aplicação *Web* a se e só se o peso da aplicação W_a for igual ao máximo dos pesos das aplicações em todas as zonas $MAX_j(W_j)$.

Se o índice do peso máximo da aplicação encontrada for igual ao índice da aplicação do *trace* de teste utilizado, então o *trace* da aplicação foi bem classificado (VP). Posteriormente, é feito o incremento dos *traces* VP, tal como dos *traces* mal classificados (FP) referentes a cada aplicação. É feito o cálculo da precisão de classificação aplicando a fórmula anteriormente mencionada na seção 3.3, dividindo o total dos *traces* VP e dos FP pelo número dos *traces* de testes referente a cada aplicação, por forma a obter as percentagens VP e FP das diversas aplicações .

$$F \in a \leftrightarrow W_a = MAX_j(W_j), j = 1, 2, ..., A$$

- F *trace* referente a uma aplicação.
- a é a aplicação ($a = 1, 2, 3, 4, 5$).
- W_a é peso da aplicação a (em todas as zonas.
- $MAX_j(W_j)$ é o peso máximo da aplicação nas zonas.
- A é o número máximo das aplicações.

De modo a avaliar a capacidade da metodologia de classificação proposta, foram realizadas várias medições de tráfego. O tráfego analisado foi recolhido com o *modo promíscuo* desactivado, capturando todo o tráfego enviado a cada cliente ligado a uma rede *wired* do laboratório de redes do Instituto de Telecomunicações. O tráfego foi capturado na direção *download*, tendo-se recolhido o número de bytes em intervalos de 300 s, como mostram as figuras 4.3, 4.4, 4.5, 4.6 , 4.7 e 4.8. A métrica da camada 2 considerada para análise foi o número de bytes capturados por intervalo de amostragem (0.1 segundos).

Foram utilizados cinco serviços *on-line* para a análise:

- **redes sociais:** o tráfego de redes sociais foi gerado usando uma conta criada no *website* Facebook (www.facebook.com), interagindo com as atualizações de notícias provenientes dos demais utilizadores ligados, o que incluía comentários e "likes";
- **notícias *on-line*:** o tráfego de notícias *on-line* foi gerado através da visita ao site do jornal Português de desporto Abola (www.abola.pt);
- **e-mail *on-line*:** o tráfego *e-email on-line* foi gerado usando os serviços oferecidos pelo serviço de *e-mail* Hotmail, mais concretamente o tráfego gerado apenas pelas sincronizações automáticas entre o terminal *web* do cliente e o servidor do Hotmail;
- **Partilha de fotos:** para gerar tráfego de uma aplicação de partilha de fotos *on-line* foi criada uma conta no *website* Flickr (www.flickr.com), tendo apenas sido considerado para análise o tráfego gerado enquanto se navegava entre as fotos de outros utilizadores;
- **serviços de vídeos:** o tráfego de *download* de vídeos *on-line* foi gerado assistindo a vídeos no YouTube.

A tabela 4.1 mostra o mapeamento entre as aplicações *web* e os serviços Internet que foram utilizados para gerar tráfego de cada serviço.

Tal como mencionado nos capítulos 1 e 3, foi efectuada uma decomposição wavelet das métricas de tráfego da camada 2 recorrendo à CWT. Os escalogramas obtidos foram normalizados por todo o comprimento do processo, como descrito na equação 3.9. As figuras

Tabela 4.1: Aplicações *On-Line* e respectivos *websites*

Serviços	**Web site**
Redes Sociais	Facebook (www.facebook.com)
Notícias *On-Line*	Abola (www.abola.com)
E-mail *On-Line*	Hotmail (www.hotmail.com)
Partilha de Fotos	Flickr (www.flickr.com)
Vídeo *On-Line*	YouTube (www.youtube.com)

de 4.3 até 4.8 mostram as métricas de tráfego, a taxa de *download* em bytes por segundo (amostrados em intervalos de 0.1 segundos) e os escalogramas correspondentes ás diferentes aplicações *web* que foram estudadas. A análise destas figuras revela caraterísticas diferentes que são causadas pelos padrões de tráfego distintos apresentados pelas aplicações, e que têm origem nas caraterísticas das interações humanas e da rede/serviços.

A aplicação de rede social *on-line* Facebook (Figura 4.3) gera tráfego que apresenta picos mais frequentes e de menor amplitude, que são gerados pelas atualizações de estado criados pelas ligações de outros utilizadores, que normalmente consistem apenas em mensagens, e também pela solicitação de uma nova página (perfil), comentários, "likes". Há portanto menos componentes de baixa frequência, ao passo que as componentes de alta frequência estão menos presentes no processo devido à pequena quantidade de tráfego trocado.

O tráfego de notícias *on-line* referente ao *site* A bola (Figura 4.4) apresenta vários picos aperiódicos de curta duração e amplitude considerável. Estes picos são causados pelos clique dos utilizadores em *hiperlinks*, enquanto navegam pelas notícias disponíveis, causando o *download* de uma nova página que apresenta as notícias solicitadas e criando componentes consideráveis de baixa frequência. Além disso, os escalogramas gerados por esta aplicação apresentam alguns componentes consideráveis de frequência intermédia, devido ao número considerável de sessões TCP criadas, existindo algumas componentes consideráveis de alta frequência relacionadas com a chegada dos pacotes.

A aplicação de *e-mail on-line*, o Hotmail, (Figura 4.5) gera tráfego com poucos picos e pouco frequentes, que correspondem à sincronização inicial e automática entre o servidor e o cliente. Estes picos são de curta duração e menos frequentes do que os das aplicações *on-line* apresentadas anteriormente. No entanto, há pequenas componentes de alta frequência causadas pelo tráfego de sincronização que raramente procura por novos *e-mails*, enquanto que as componentes de baixa frequência não estão muitos espalhadas pelo escalograma de tráfego devido à natureza periódica dos eventos de rede/serviços.

A aplicação de partilha de fotos *on-line* Flickr (Figura 4.6) gera vários picos de tráfego com pseudo-periodicidade, devido aos cliques que são executados pelo utilizador enquanto pede para ver outra fotografia. Estes picos são geralmente de baixa amplitude, já que consistem apenas no *download* de uma fotografia utilizando uma única sessão TCP. Consequentemente, podem-se observar várias componentes de alta frequência e de baixa amplitude, espalhadas por todo o escalograma, existindo também alguns componentes de baixa frequência.

Por último, o serviço de partilha de vídeos *on-line* Youtube (figuras 4.7 e 4.8) gera tráfego com intervalos pequenos entres chegadas de pacotes, causados pelo download do vídeo solicitado na largura de banda da rede que neste caso está totalmente disponível. Consequentemente, existem componentes consideráveis de alta frequência causadas pela chegada dos pacotes, não existindo componentes relevantes de baixa frequência já que não há muitos cliques por parte dos utilizadores.

Facebook

A escolha deste *site* advém do crescente número de utilizadores que utilizam o Facebook para as suas várias tarefas do dia-a-dia.

O tráfego de redes sociais foi gerado utilizando uma conta criada no Facebook (registo feito com uma conta prévia no site Hotmail) e apresenta picos frequentes (ver figura 4.3) que são gerados pelas atualizações de estado criadas por outros utilizadores que também estão ligados, que normalmente consistem apenas em mensagens, e também pela solicitação de uma nova página (perfil), comentários, "likes", etc.

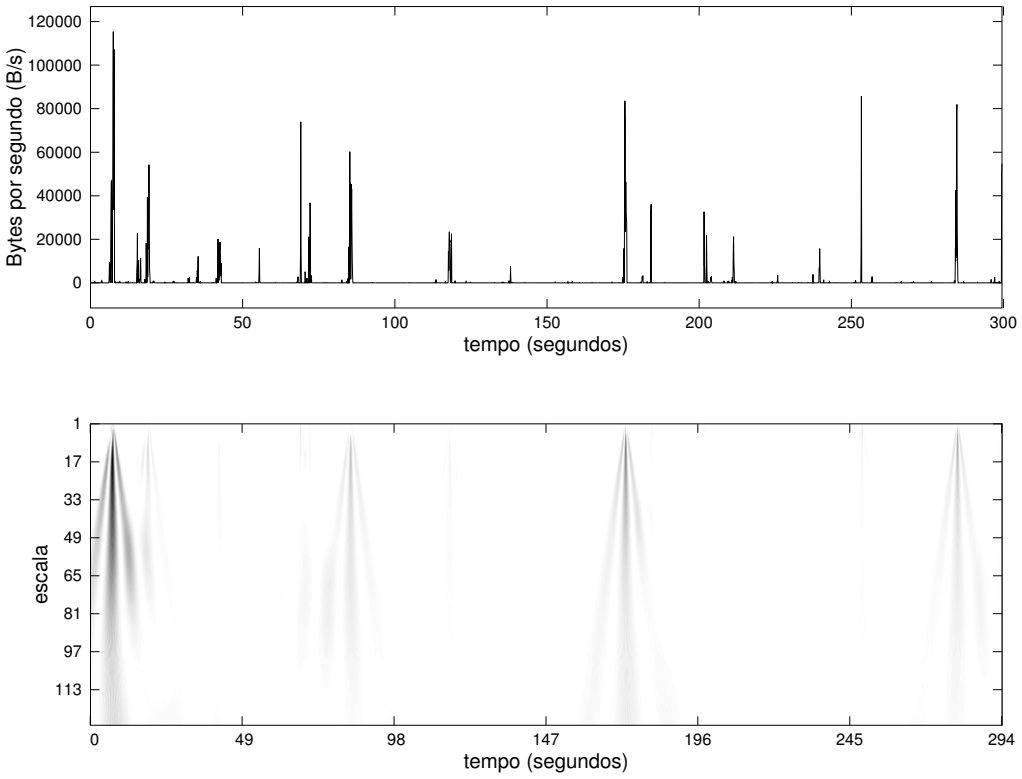


Figura 4.3: Padrões de Tráfego da Rede Social *On-line* Facebook - direção download e Escalograma correspondente.

Abola

Tal como o Facebook, o site Abola foi escolhido devido à sua popularidade. O tráfego da Abola (figura 4.4) apresenta vários picos aperiódicos de curta duração. Esses picos são causados pelo utilizador sempre que este clica em *hiperlinks* enquanto navega através das notícias disponíveis, fazendo *download* de uma nova página que apresenta a notícia solicitada.

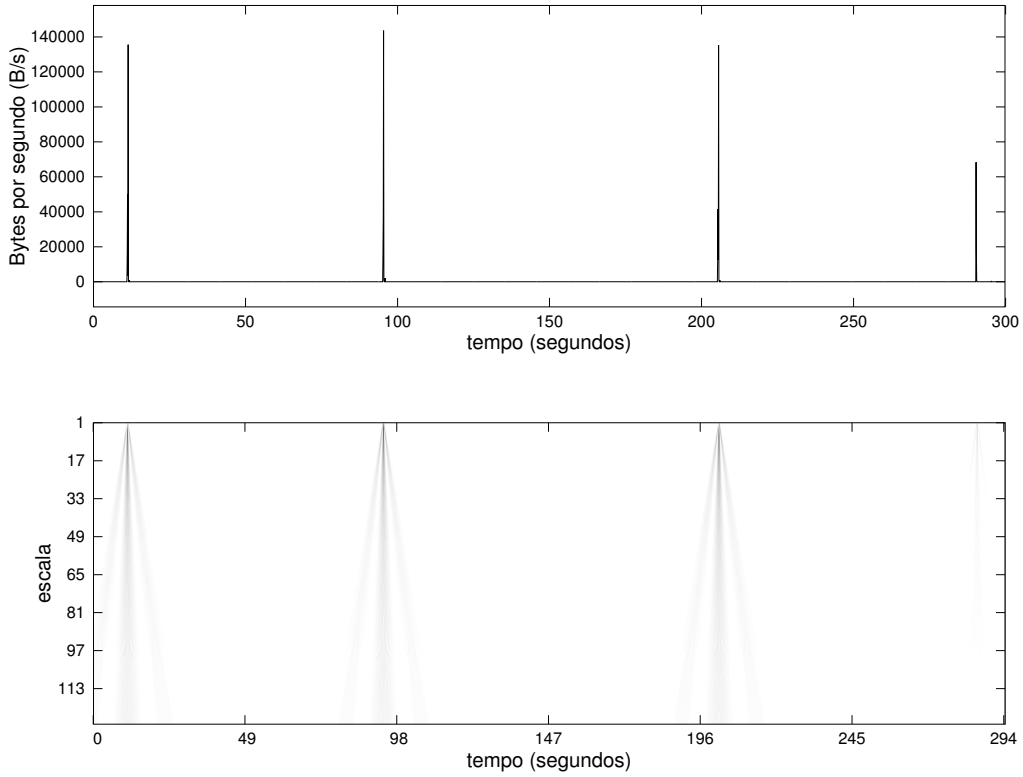


Figura 4.4: Padrões de Tráfego do Serviço de Notícias *On-line* Abola - direção download e Escalograma correspondente.

Hotmail

A escolha deste *site* também se deveu à popularidade, pois é considerado um dos maiores serviços de *e-mail* no mundo. O Hotmail conta com mais de 360 milhões de utilizadores.

O tráfego gerado pelo Hotmail, como mostra a figura 4.5, apresenta poucos picos e pouco frequentes, que correspondem ás sincronizações inicial e automáticas entre o terminal *web* do cliente e o servidor do Hotmail. Estes picos têm duração muito curta e são pouco frequentes porque o tráfego de sincronização apenas verifica a existência de novos *e-mails*.

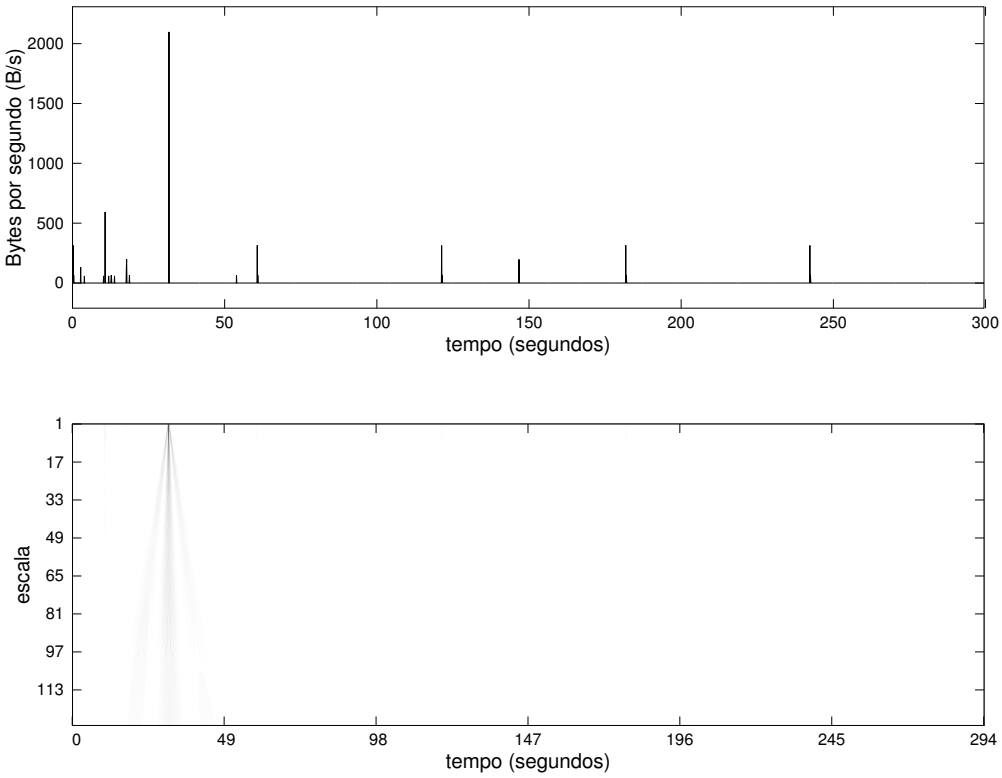


Figura 4.5: Padrões de Tráfego dos serviços de *E-mails Online* Hotmail - direção download e Escalograma correspondente .

Flickr

Para gerar tráfego, foi criada uma conta no Flickr. Note-se que o registo no *site* só é possível caso se tenha feito um registo no Yahoo. Foi considerado para análise o tráfego gerado enquanto se navegava em fotos de outros utilizadores.

A aplicação de partilha de fotos (Flickr) gera vários picos de tráfego com pseudo - periodicidade, como se pode observar na figura 4.6, devido aos cliques que são executados pelo utilizador enquanto pede para ver outra fotografia.

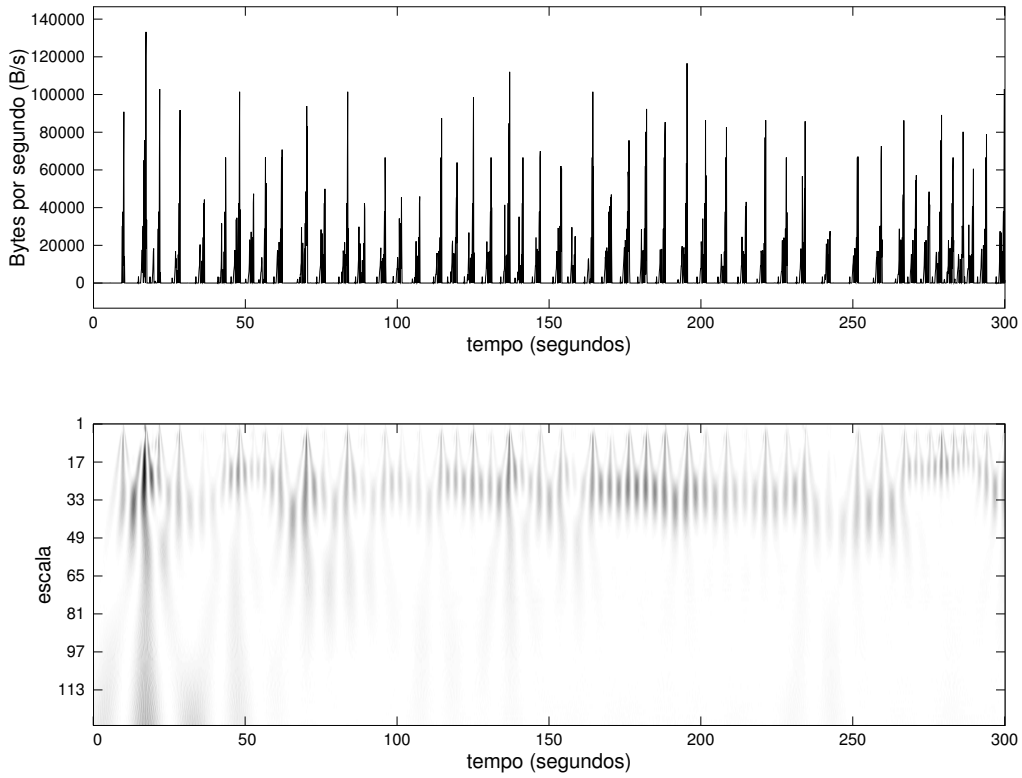


Figura 4.6: Padrões de Tráfego de Serviço de Partilha de Fotos Flickr - direção download e Escalograma correspondente.

Youtube

O tráfego de *download* de vídeo *on-line* foi gerado assistindo a vídeos em HD. Este serviço gera tráfego com alguns picos e com alguma frequência, como mostram as figuras 4.7 e 4.8, uma vez que o número de cliques do utilizador não é tão relevante.

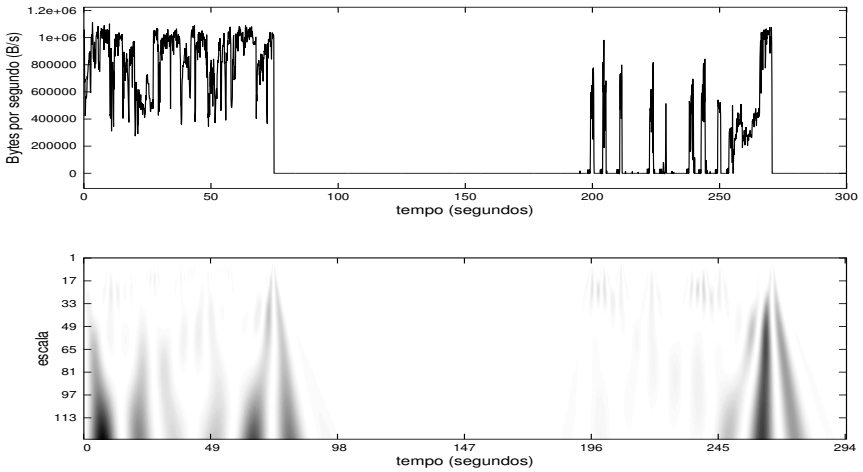


Figura 4.7: Padrões de Tráfego de Serviço de Partilha de Vídeos Youtube (VCD) - direção download e Escalograma correspondente.

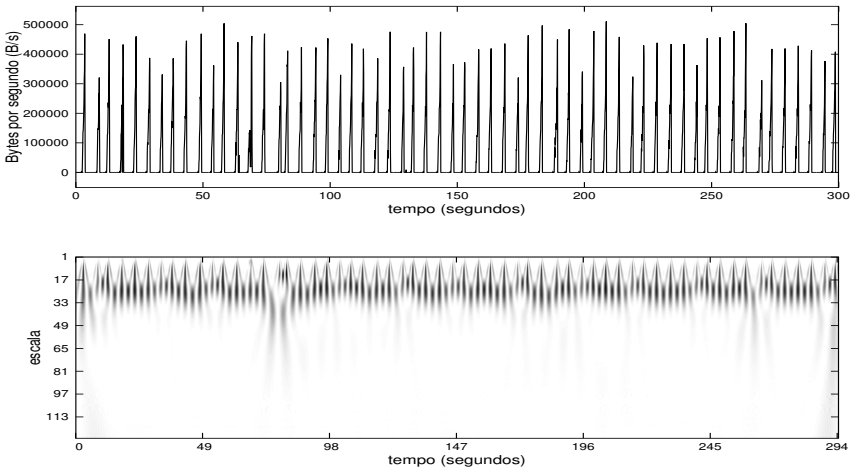


Figura 4.8: Padrões de Tráfego de Serviço de Partilha de Vídeos (VLD) - direção download e Escalograma correspondente.

4.3 Sumário

Neste capítulo foi apresentada a metodologia de identificação das diferentes aplicações *web*. Aplicando uma CWT sobre as métricas de tráfego da camada 2, obtemos os escalogramas de tráfego das diferentes aplicações *web*. Efectuando a análise dos componentes de frequência destes escalogramas é possível identificar as regiões do espetro de frequência correspondentes aos eventos característicos de cada aplicação *web*. Através do cálculo do desvio padrão dos coeficientes Wavelet obtemos os comportamentos dos *traces* das varias aplicações e calculando o peso destes *traces* ao longo das regiões é possível identificar o tráfego gerado pelos diferentes serviços de Internet estudados.

Capítulo 5

Apresentação e discussão de resultados

Neste capítulo será efectuada a análise e a discussão dos resultados obtidos, de forma a identificarmos a melhor metodologia de classificação de tráfego.

A discussão dos resultados é feita através da análise das tabelas obtidas a partir da *matriz de classificação* (seção 3.3). A comparação das tabelas para o mesmo *site* e para *sites* diferentes facilita a compreensão do que realmente aconteceu e a discussão de resultados.

5.1 Resultados

A partir da *matriz de classificação*, definida na seção 3.3, obtemos as tabelas 5.2 e 5.3 correspondentes aos algoritmos 1 e 2, respectivamente, onde é possível observar os resultados detalhados da classificação de todas as aplicações analisadas.

Foram consideradas 5 aplicações ($A = 5$), tendo-se subdividido a aplicação de partilha de vídeos em dois tipos: Vídeos de Longa Duração VLD e Vídeos de Curta Duração VCD. Para a validação dos algoritmos de classificação foram utilizados no total 360 *traces*, 60 por cada aplicação, exceptuando a aplicação de partilha de vídeos que contou com 120 *traces*, 60 para VLD e 60 para VCD .

Para calcular o Intervalo de Confiança (IC) foram efetuadas cinco iterações; em cada iteração foram escolhidos aleatoriamente doze *traces* de treino e doze *traces* de teste por aplicação. Para o varrimento dos pontos dos *traces* ao longo das regiões foram definidos os valores 8, 0.0003, 1, 8 e 0.0003 para as variáveis w_x, w_y, α, d_x e d_y, respetivamente (seção 4.1).

O facto de efetuar varias iterações permite minimizar o fator de imprecisão e variação dos resultados obtidos nas classificações, tornando portanto a nossa abordagem mais confiável. Para cada algoritmo o resultado apresentado é a média dos resultados obtidos nas cinco iterações.

A tabela 5.1 mostra a variação dos resultados de classificação do *trace* da aplicação de rede social Facebook, onde é possível identificar a volubilidade da precisão de classificação mediante a variação das variáveis w_x, w_y, α, d_x e d_y.

Tabela 5.1: Tabela de classificação Facebook.

Resultado da classificação					
Aplicações Web		**Classificação variaveis**			
		w_x=8 w_y=0.0003 α=0.5 d_x=α*w_x=4 d_y=α*w_y=0.00015	w_x=8 w_y=0.0003 α=1 d_x=α*w_x=8 d_y=α*w_y=0.0003	w_x=16 w_y=0.0006 α=0.5 d_x=α*w_x=8 d_y=α*w_y=0.0003	w_x=16 w_y=0.0006 α=1 d_x=α*w_x=16 d_y=α*w_y=0.0006
facebook	Algoritmo 1	50%	51.67%	58.33%	25%
	Algoritmo 2	91.67%	83.33%	66.67%	91.67%

Seguidamente são apresentadas as tabelas 5.2 e 5.3 com os resultados obtidos por aplicação da nossa abordagem, utilizando variáveis w_x, w_y, α, d_x e d_y com valores iguais a 8, 0.0003, 1, 8, 0.0003 respetivamente, por serem estes valores que conduzem a resultados que em média satisfazem os requisitos anteriormente apresentados, ou seja, garantem a obtenção de uma melhor precisão de classificação para cada aplicação *Web* estudada.

Tabela 5.2: Tabela de classificação Algoritmo 1.

Resultado da classificação (%)							
Aplicações Web		**Classificação**					
		facebook	**abola**	**hotmail**	**flickr**	**youtube**	
						LD	**CD**
facebook		**51,67**	26,66	15,00	5,00	1,67	0
	IC	35,83-67,51	17,15-36,19	8,89-21,11	0-11,53	0-4,93	0-0
abola		0	**76,67**	21,66	1,67	0	0
	IC	0-0	64,66-88,67	10,58-32,74	0-4,93	0-0	0-0
hotmail		3,33	1,67	**90,00**	5,00	0	0
	IC	0-7,33	0-4,93	83,89-96,11	1-9	0-0	0-0
flickr		1,67	1,67	1,67	**91,65**	1,67	1,67
	IC	0-4,93	0-4,93	0-4,93	86,50-96,83	0-4,93	0-4,93
Youtube	**LD**	3,33	0	3,33	23,33	**70,01**	0
	IC	0-7,33	0-0	0-7,33	9,27-37,38	56,10-84,57	0-0
	CD	0	0	0	1,67	0	**98,33**
	IC	0-0	0-0	0-0	0-4,93	0-0	95,06-100,00

Tabela 5.3: Tabela de classificação Algoritmo 2

Resultado da classificação (%)							
Aplicações Web		**Classificação**					
		facebook	**abola**	**hotmail**	**flickr**	**youtube**	
						LD	**CD**
facebook		**83,33**	11,67	3,33	1,67	0	0
	IC	76,03-90,64	7,66-15,67	0-7,33	0-4,9313	0-0	0-0%
abola		25,00	**63,33**	11,67	0	0	0
	IC	10,39-39,61	48,19-78,48	0-27,67	0-0	0-0	0-0
hotmail		13,33	1,67	**81,67**	0	3,33	0
	IC	3,53-23,13	0-4,93	69,67-93,67	0-0	0-9,86	0-0
flickr		25,00	1,67	0	**70,00**	3,33	0
	IC	13,45-36,55	0-4,93	0-0	58,92-81,08	0-7,32	0-0
Youtube	**LD**	41,67	0	0	3,33	**55,00**	0
	IC	34,36-48,97	0-0	0-0	0-7,33	45,2-64,8	0-0
	CD	18,33	0	0	0	0	**81,67**
	IC	2,49-34,17	0-0	0-0	0-0	0-0	65,83-97,51

Analisemos agora os resultados de classificação que foram obtidos através dos métodos apresentados no capítulo 4 e recorrendo aos dados apresentados das tabelas 5.2 e 5.3. Pode-se afirmar que a maior parte do tráfego gerado é mapeado com precisão na aplicação *web* correspondente. No entanto, existem alguns erros de classificação que podem ser explicados.

Alguns *traces* de redes sociais foram associados a notícias *on-line*, o que pode acontecer se um número considerável de atualizações do estado (notificações recebidas no *feed* de notícia da rede social) ocorrer num curto período de tempo.

As funcionalidades comuns de *e-mail* e troca de mensagens são todas evidenciadas pelo facebook, como a partilha de fotos e vídeos e a possibilidade de atualizar o 'estado atual' (o que o utilizador está a fazer ou como se sente) a partir da página do seu perfil, criando assim fluxos que se assemelham aos serviços de partilha de fotos (Flickr), partilha de vídeos (Youtube) e de *e-mail*.

A classificação de alguns *traces* de serviços de redes sociais *on-line* como pertencentes a serviços de partilha de fotos pode ocorrer quando um utilizador visita o perfil de outro utilizador ligado à rede, visualizando fotos, o que se traduz num comportamento que se assemelha à visualização de fotos no flickr.

Alguns *traces* de aplicações de *e-mail* foram igualmente associados a aplicações de redes sociais, o que pode ser explicado pelo fato de que quando existe uma pequena quantidade de atualizações de *e-mails* o perfil da aplicação se torna mais semelhante à troca de mensagens de redes sociais.

Alguns fluxos de aplicações de *e-mail* foram atribuídos a notícias *on-line*, devido ás notificações recebidas automaticamente pela aplicação.

Fluxos de aplicações de notícias *on-line* podem ser associados a serviços de partilha de fotos, isto porque os serviços de notícias normalmente contêm imagens embutidas.

As aplicações de redes sociais, notícias *on-line* e serviços de *e-mail* apresentam igualmente algumas associações entre si (ou seja, fluxos de uma aplicação que são erradamente classificados como pertencentes a outra), isto porque estas são aplicações dinâmicas que recebem

atualizações automáticas, fazendo com que tenham fluxos que apresentam características semelhantes.

Alguns erros de classificação também ocorrem nas aplicações de partilha de fotos, onde alguns *traces* foram classificados como pertencentes a redes de partilha de vídeos. Isto pode ser explicado pelo fato da aplicação de partilha de fotos apresentar algumas páginas com vídeos.

Quanto à classificação do serviço de partilha de vídeos, a periodicidade do *player* do Youtube pode assemelhar-se a eventos realizados pelo utilizador quando efetua cliques periódicos para visualizar novas fotografias (flickr,facebook) ou solicitar novas páginas. Isto pode levar a que alguns fluxos de vídeos sejam associados a aplicações de partilha de fotos e redes sociais.

Quanto à classificação dos serviços, podemos afirmar que o algoritmo 1 apresenta melhores resultados, com exceção do serviço de rede social (Facebook) que apresenta melhor desempenho com o algoritmo 2.

É possível constatar que, em média, o algoritmo 1 obteve uma melhor precisão na classificação (melhor desempenho). De facto, para os vários *traces* das diversas aplicações o algoritmo 1 apresenta um maior taxa de acertos, ou seja, a taxa de *traces* de aplicações VP é maior quando comparado com o algoritmo 2.

A diferença entre os algoritmos, nesse caso, não é significativa, mas indica que o algoritmo 1 apresenta um alto grau de precisão.

5.2 Sumário

Neste capítulo foram apresentados e analisados os resultados obtidos utilizando os dois algoritmos de classificação propostos. A metodologia de classificação foi aplicada aos *traces* de tráfego das aplicações estudadas. Esses *traces* foram passivamente recolhidos na rede da Universidade de Aveiro, tendo sido medidos entre Outubro de 2011 e Setembro de 2012. Foram consideradas 5 aplicações ($A = 5$), tendo-se ainda subdividido a aplicação de partilha de vídeos em dois tipos: VLD e VCD. Os resultados da classificação, apesar de satisfatórios, apresentam alguns erros: certos *traces* de uma determinada aplicação *web* foram associados a outras aplicações, devido ao facto de existirem semelhanças nos serviços utilizados.

Capítulo 6

Conclusão

O aparecimento da Internet como plataforma de comunicação levou à necessidade de criar mecanismos capazes de caracterizar com precisão cada utilizador e o tráfego por ele gerado.

Com a crescente expansão dos serviços, aplicações e do número de utilizadores da Internet surgem novos desafios para os investigadores, ISPs e para os próprios utilizadores. A crescente concorrência no mercado leva a que os ISPs aumentem constantemente a capacidade dos serviços oferecidos, dando origem a graves problemas de gestão de rede. Além disso, o crescimento exponencial das aplicações de Internet, associado ao crescente número de solicitações, despoletou a necessidade de um mapeamento preciso do tráfego de Internet com a aplicação que lhe deu origem.

Outra vantagem do desenvolvimento de metodologias de classificação de tráfego advém da necessidade da realização de uma gestão eficiente da infra-estrutura da rede, assim como da necessidade de identificar atempadamente tráfego ilícito e suspeito.

A crescente complexidade da Internet, onde existem várias politicas de privacidade que limitam a aplicabilidade de determinadas técnicas de análise do conteúdo do tráfego trocado, é uma das várias razões para se criarem novas metodologias de classificação que consigam lidar com essas restrições.

Esta dissertação aborda esta questão, propondo uma abordagem de classificação adequada que possa ser implementada em cenários com diversas restrições.

Outro problema recente e que tem crescido de uma forma alarmante são os ataques de segurança na Internet: a identificação atempada do tráfego que apresenta padrões ilícitos e/ou suspeitos pode ter uma importância vital na prevenção desses ataques ou na limitação das suas consequências.

O principal conceito por trás da abordagem proposta consiste na análise do tráfego de diferentes aplicações *web*, tendo em especial atenção os eventos e mecanismos por elas gerados.

Verificou-se que diferentes aplicações da Internet requerem diferentes interações por parte do utilizador, e que as solicitações por ele afetuadas através de uma página *web* criam um conjunto de sessões de Internet que por sua vez criam um conjunto de pacotes que são transmitidos através da ligação física. Estes eventos criam várias componentes de frequência em diferentes regiões do espetro. Os eventos podem ser classificados em três tipos: eventos que apresentam componentes de baixa frequência constituem eventos humanos, associados aos comportamentos e ações humano/utilizador, como cliques de um utilizador numa página web; eventos com componentes de frequência média, onde se incluem as sessões de tráfego e os mecanismos de controlo do tráfego; e por último, eventos que apresentam componentes de

alta frequência, correspondentes aos eventos dos protocolos e da Internet, tais como chegadas de pacotes.

A nossa abordagem permite identificar as componentes de frequência presentes nos *data-streams* através de uma decomposição wavelet das métricas do tráfego capturado, permitindo efetuar uma analise multi-escalar de cada aplicação Internet estudada. Através do cálculo do desvio padrão dos coeficientes Wavelet obtemos os comportamentos dos *traces* das várias aplicações estudadas; calculando o peso destes traces ao longo das regiões é possível classificar o tráfego gerado pelos diferentes serviços de Internet que foram considerados.

A abordagem desenvolvida permite um classificação precisa e eficiente do tráfego gerado pelas aplicações de Internet. No entanto alguns *traces* de tráfego utilizados foram mal classificados, não sendo possível obter um mapeamento totalmente preciso do tráfego de algumas aplicações *web*. Esta desvantagem advém da associação de um *trace* de uma determinada aplicação com outras aplicações, o que ocorre devido ao facto de existirem semelhanças de comportamentos entre as diversas aplicações estudadas.

6.1 Sugestões para trabalhos futuros

Como trabalho futuro, podemos referir o melhoramento na precisão da classificação utilizando mais métricas de tráfego no processo de identificação das aplicações *web*. Outra tarefa poderá passar pelo desenvolvimento de melhores plataformas de monitorização, amostragem e análise de tráfego para a nossa abordagem de classificação, bem como a utilização de abordagens eficientes para analisar componentes presentes nas diferentes regiões do espetro de frequência. Note-se que nesta dissertação apenas foi caraterizado o tráfego de aplicações *web*, podendo o desenvolvimento de técnicas para profiling de utilizadores ser outra tarefa para trabalho futuro. O desenvolvimento de mais algoritmos de classificação baseados nos pesos dos *traces* das aplicações ao longo das regiões é outro trabalho que deve ser desenvolvido no futuro.

Bibliografia

[1] I. T. U. (ITU). //http://www.itu.int/ITU-D/ict/statistics/, - acedido em Agosto 2012.

[2] C. V. G. Forecast, "Visual networking index (vni)." //http://www.cisco.com/web/PT/press/articles/2012/20120601.html, - acedido em Agosto 2011.

[3] E. Rocha, "Metodologias para caracterização de tráfego em redes de comunicação," in *Tese de Doutoramento*, Universidade de Aveiro, 2011.

[4] B. Grobauer, T. Walloschek, and E. Stocker, "Understanding cloud computing vulnerabilities," *Security Privacy, IEEE*, vol. 9, pp. 50 –57, march-april 2011.

[5] I. Drago and A. Pras, "Scalable service performance monitoring," in *Proceedings of the Mechanisms for autonomous management of networks and services, and 4th international conference on Autonomous infrastructure, management and security*, AIMS'10, (Berlin, Heidelberg), pp. 175–178, Springer-Verlag, 2010.

[6] F. McSherry and R. Mahajan, "Differentially-private network trace analysis," in *Proceedings of the ACM SIGCOMM 2010 conference*, SIGCOMM '10, (New York, NY, USA), pp. 123–134, ACM, 2010.

[7] A. Callado, C. Kamienski, G. Szabo, B. Gero, J. Kelner, S. Fernandes, and D. Sadok, "A survey on internet traffic identification," *Communications Surveys Tutorials, IEEE*, vol. 11, pp. 37 –52, quarter 2009.

[8] P. numbers, July 2012. `http://www.iana.org/assignments/port-numbers`.

[9] D. Moore, K. Keys, R. Koga, E. Lagache, and K. C. Claffy, "The CoralReef software suite as a tool for system and network administrators," in *LISA '01: Proceedings of the 15th USENIX conference on System administration*, (Berkeley, CA, USA), pp. 133–144, USENIX Association, 2001.

[10] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-service mapping for qos: a statistical signature-based approach to ip traffic classification," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, IMC '04, (New York, NY, USA), pp. 135–148, ACM, 2004.

[11] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in-network identification of p2p traffic using application signatures," in *WWW '04: Proceedings of the 13th international conference on World Wide Web*, (New York, NY, USA), pp. 512–521, ACM, 2004.

[12] A. Madhukar and C. Williamson, "A longitudinal study of p2p traffic classification," in *Proceedings of the 14th IEEE International Symposium on Modeling, Analysis, and Simulation*, MASCOTS '06, (Washington, DC, USA), pp. 179–188, IEEE Computer Society, 2006.

[13] T. Karagiannis, A. Broido, N. Brownlee, K. Claffy, and M. Faloutsos, "File-sharing in the internet: A characterization of P2P traffic in the backbone," tech. rep., University of California, Riverside, Riverside, CA, USA, Nov. 2003.

[14] T. Karagiannis, A. Broido, N. Brownlee, k. claffy, and M. Faloutsos, "Is p2p dying or just hiding?," in *Global Internet and Next Generation Networks*, (Dallas, Texas), Globecom 2004, Dec 2004.

[15] A. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," in *Passive and Active Network Measurement* (C. Dovrolis, ed.), vol. 3431 of *Lecture Notes in Computer Science*, pp. 41–54, Springer Berlin / Heidelberg, 2005.

[16] C. Dewes, A. Wichmann, and A. Feldmann, "An analysis of internet chat systems," in *IMC '03: Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, (New York, NY, USA), pp. 51–64, ACM Press, 2003.

[17] P. Haffner, S. Sen, O. Spatscheck, and D. Wang, "Acas: Automated construction of application signatures," in *In SIGCOMMâ05 MineNet Workshop*, 2005.

[18] T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *Communications Surveys Tutorials, IEEE*, vol. 10, pp. 56 –76, quarter 2008.

[19] A. W. Moore and D. Zuev, "Internet traffic classification using bayesian analysis techniques," in *Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, vol. 33 of *SIGMETRICS '05*, (New York, NY, USA), pp. 50–60, ACM, June 2005.

[20] S. Zander, T. T. T. Nguyen, and G. J. Armitage, "Automated traffic classification and application identification using machine learning.," in *LCN*, pp. 250–257, IEEE Computer Society, 2005.

[21] A. McGregor, M. Hall, P. Lorier, and J. Brunskill, "Flow clustering using machine learning techniques.," in *PAM* (C. Barakat and I. Pratt, eds.), vol. 3015 of *Lecture Notes in Computer Science*, pp. 205–214, Springer, 2004.

[22] Y. Hu, D.-M. Chiu, and J. C. S. Lui, "Application identification based on network behavioral profiles.," in *IWQoS* (H. van den Berg and G. Karlsson, eds.), pp. 219–228, IEEE, 2008.

[23] T. Karagiannis, A. Broido, M. Faloutsos, and K. Claffy, "Transport layer identification of P2P traffic," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, IMC '04, (New York, NY, USA), pp. 121–134, ACM, 2004.

[24] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification," in *In IMC'04*, pp. 135–148, 2004.

[25] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: multilevel traffic classification in the dark," in *Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications*, vol. 35 of *SIGCOMM '05*, (New York, NY, USA), pp. 229–240, ACM, 2005.

[26] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, "Traffic classification on the fly," *SIGCOMM Comput. Commun. Rev.*, vol. 36, pp. 23–26, Apr. 2006.

[27] J. Erman, M. Arlitt, and A. Mahanti, "Traffic classification using clustering algorithms," in *Proceedings of the 2006 SIGCOMM workshop on Mining network data*, MineNet '06, (New York, NY, USA), pp. 281–286, ACM, 2006.

[28] Y. Hu, D.-M. Chiu, and J. C. S. Lui, "Profiling and identification of p2p traffic," *Comput. Netw.*, vol. 53, pp. 849–863, Apr. 2009.

[29] J. Hurley, E. Garcia-Palacios, and S. Sezer, "Classifying network protocols: A 'two-way' flow approach," *Communications, IET*, vol. 5, pp. 79 –89, january 2011.

[30] J. Gao, G. Hu, X. Yao, and R. Chang, "Anomaly detection of network traffic based on wavelet packet," in *Communications, 2006. APCC '06. Asia-Pacific Conference on*, pp. 1 –5, 31 2006-sept. 1 2006.

[31] W. Lu and A. A. Ghorbani, "Network anomaly detection based on wavelet analysis.," *EURASIP J. Adv. Sig. Proc.*, vol. 2009, 2009.

[32] T. Nguyen and G. Armitage, "Training on multiple sub-flows to optimise the use of machine learning classifiers in real-world ip networks," in *Local Computer Networks, Proceedings 2006 31st IEEE Conference on*, pp. 369 –376, nov. 2006.

[33] D. Bonfiglio, M. Mellia, M. Meo, D. Rossi, and P. Tofanelli, "Revealing skype traffic: when randomness plays with you," in *Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '07, (New York, NY, USA), pp. 37–48, ACM, 2007.

[34] N.-F. Huang, G.-Y. Jai, and H.-C. Chao, "Early identifying application traffic with application characteristics," in *Communications, 2008. ICC '08. IEEE International Conference on*, pp. 5788 –5792, may 2008.

[35] "Weka 3 - data mining with open source machine learning software in java."

[36] W. Jiang and M. Gokhale, "Real-time classification of multimedia traffic using fpga," in *Field Programmable Logic and Applications (FPL), 2010 International Conference on*, pp. 56 –63, 31 2010-sept. 2 2010.

[37] J. But, P. Branch, and T. Le, "Rapid identification of bittorrent traffic," in *Local Computer Networks (LCN), 2010 IEEE 35th Conference on*, pp. 536 –543, oct. 2010.

[38] D. Godoy and A. Amandi, "User profiling in personal information agents: a survey," *Knowl. Eng. Rev.*, vol. 20, pp. 329–361, Dec. 2005.

[39] J. Iglesias, P. Angelov, A. Ledezma, and A. Sanchis, "Creating evolving user behavior profiles automatically," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, pp. 854 –867, may 2012.

[40] G. Adomavicius and A. Tuzhilin, "User profiling in personalization applications through rule discovery and validation," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, (New York, NY, USA), pp. 377–381, ACM, 1999.

[41] S. N. Schiaffino and A. Amandi, "Intelligent user profiling.," in *Artificial Intelligence: An International Perspective* (M. Bramer, ed.), vol. 5640 of *Lecture Notes in Computer Science*, pp. 193–216, Springer, 2009.

[42] Y. Moreau, P. B. J. Shawe-taylor, C. Stoermann, S. Ag, and C. C. Vodafone, "Novel techniques for fraud detection in mobile telecommunication networks," 1996.

[43] R. Buschkes, D. Kesdogan, and P. Reichl, "How to increase security in mobile networks by anomaly detection," in *Computer Security Applications Conference, 1998. Proceedings. 14th Annual*, pp. 3 –12, dec 1998.

[44] T. Fawcett and F. J. Provost, "Adaptive fraud detection," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 291–316, 1997.

[45] S. H. Oh and W. S. Lee, "An anomaly intrusion detection method by clustering normal user behavior," *Computers and Security*, pp. 596–612, 2003.

[46] C. Manikopoulos and S. Papavassiliou, "Network intrusion and fault detection: a statistical anomaly approach," *Communications Magazine, IEEE*, vol. 40, pp. 76 – 82, oct 2002.

[47] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," *SIGKDD Explor. Newsl.*, no. 1, pp. 50–59.

[48] P. Mcdaniel, J. V. D. Merwe, S. Sen, B. Aiello, O. Spatscheck, and C. Kalmanek, "Enterprise security: a community of interest based approach," in *In Proc. NDSS*, 2006.

[49] H. joon Kim, S. Lee, B. Lee, and S. Kang, "Building concept network-based user profile for personalized web search," in *Computer and Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference on*, pp. 567 –572, aug. 2010.

[50] opennms, "The open nms project." http://www.opennms.org/, August 2012.

[51] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Googling the internet: Profiling internet endpoints via the world wide web," *Networking, IEEE/ACM Transactions on*, vol. 18, pp. 666 –679, april 2010.

[52] I. Trestian, S. Ranjan, A. Kuzmanovi, and A. Nucci, "Unconstrained endpoint profiling (googling the internet)," in *Proceedings of the ACM SIGCOMM 2008 conference on Data communication*, SIGCOMM '08, (New York, NY, USA), pp. 279–290, ACM, 2008.

[53] E. Rocha, P. Salvador, and A. Nogueira, "Classification of hidden users' profiles in wireless communications," in *MONAMI*, pp. 3–16, 2011.

[54] A. Feldmann, A. C. Gilbert, and W. Willinger, "Data networks as cascades: Investigating the multifractal nature of internet WAN traffic," pp. 42–55, 1998.

[55] T. W. Foundation. //http://www.wireshark.org/, - acedido em Agosto 2012.

[56] P. A. Morettin, "From fourier to wavelet analysis of time series," in *In Proceedings in Computational Statistics*, pp. 111–122, Physica-Verlag, 1996.

[57] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bull. Amer. Meteor. Soc.*, vol. 79, pp. 61–78, Jan. 1998.

[58] D. T. L. Lee and A. Yamamoto, "Wavelet analysis: Theory and applications," vol. 45, pp. 44–54, Dec. 1994.

[59] I. Daubechies, *Ten lectures on wavelets*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1992.

[60] J. Slavic, I. Simonovski, and M. Boltezar, "Damping identification using a continuous wavelet transform: application to real data," *Journal of Sound and Vibration*, vol. 262, no. 2, pp. 291 – 307, 2003.

[61] P. Abry, P. Flandrin, M. S. Taqqu, and D. Veitch, "Wavelets for the analysis, estimation, and synthesis of scaling data," in *Self-Similar Network Traffic and Performance Evaluation* (K. Park and W. Willinger, eds.), pp. 39–88, Wiley, 2000.

[62] K. Gurley and A. Kareem, "Applications of wavelet transforms in earthquake, wind and ocean engineering," *Doctoralś Thesis*, 1999.

[63] T. public repository. //http://www.tcpdump.org/, - acedido em Agosto 2012.

[64] WinPcap. //http://www.winpcap.org/windump/, - acedido em Agosto 2012.

[65] W. distribution. //http://www.wireshark.org/docs/man-pages/tshark.html, - acedido em Agosto 2012.

[66] G. Octave. //http://www.gnu.org/software/octave/, - acedido em Agosto 2012.

[67] M. Matlab. //http://www.mathworks.com/products/matlab/, - acedido em Agosto 2012.